Pay Attention to Features, Transfer Learn Faster CNNs
Kafeng Wang,Xitong Gao,Yiren Zhao,Xingjian Li,Dejing Dou,Cheng-Zhong Xu
Deep convolutional neural networks are now widely deployed in vision application s, but a limited size of training data can restrict their task performance. Tran sfer learning offers the chance for CNNs to learn with limited data samples by t ransferring knowledge from models pretrained on large datasets. Blindly transfer ring all learned features from the source dataset, however, brings unnecessary c omputation to CNNs on the target task. In this paper, we propose attentive featu re distillation and selection (AFDS), which not only adjusts the strength of tra nsfer learning regularization but also dynamically determines the important feat ures to transfer. By deploying AFDS on ResNet-101, we achieved a state-of-the-ar t computation reduction at the same accuracy budget, outperforming all existing transfer learning methods. With a 10x MACs reduction budget, a ResNet-101 equipp ed with AFDS transfer learned from ImageNet to Stanford Dogs 120, can achieve an accuracy 11.07% higher than its best competitor.
****

Differentiable Hebbian Consolidation for Continual Learning
Vithursan Thangarasa,Thomas Miconi,Graham W. Taylor
Continual learning is the problem of sequentially learning new tasks or knowledg e while protecting previously acquired knowledge. However, catastrophic forgetti ng poses a grand challenge for neural networks performing such learning process. Thus, neural networks that are deployed in the real world often struggle in sce narios where the data distribution is non-stationary (concept drift), imbalanced , or not always fully available, i.e., rare edge cases. We propose a Differentia ble Hebbian Consolidation model which is composed of a Differentiable Hebbian Pl asticity (DHP) Softmax layer that adds a rapid learning plastic component (compr essed episodic memory) to the fixed (slow changing) parameters of the softmax ou tput layer; enabling learned representations to be retained for a longer timesca le. We demonstrate the flexibility of our method by integrating well-known task- specific synaptic consolidation methods to penalize changes in the slow weights that are important for each target task. We evaluate our approach on the Permute d MNIST, Split MNIST and Vision Datasets Mixture benchmarks, and introduce an im balanced variant of Permuted MNIST --- a dataset that combines the challenges of class imbalance and concept drift. Our proposed model requires no additional hy perparameters and outperforms comparable baselines by reducing forgetting.
****

Generative Hierarchical Models for Parts, Objects, and Scenes
Fei Deng,Zhuo Zhi,Sungjin Ahn
Hierarchical structure such as part-whole relationship in objects and scenes are the most inherent structure in natural scenes. Learning such representation via unsupervised learning can provide various benefits such as interpretability, co mpositionality, and transferability, which are important in many downstream task s. In this paper, we propose the first hierarchical generative model for learnin g multiple latent part-whole relationships in a scene. During inference, taking top-down approach, our model infers the representation of more abstract concept (e.g., objects) and then infers that of more specific concepts (e.g., parts) by conditioning on the corresponding abstract concept. This makes the model avoid a difficult problem of routing between parts and whole. In experiments on images containing multiple objects with different shapes and part compositions, we demo nstrate that our model can learn the latent hierarchical structure between parts and wholes and generate imaginary scenes.
****

Mixture Distributions for Scalable Bayesian Inference
Pranav Poduval,Hrushikesh Loya,Rajat Patel,Sumit Jain
Bayesian Neural Networks (BNNs) provides a mathematically grounded framework to quantify uncertainty. However BNNs are computationally inefficient,
thus are generally not employed on complicated machine learning tasks. Deep Ensembles were introduced as a Bootstrap inspired frequentist approach to the community, as an alternative to BNN’s. Ensembles of deterministic and stochastic networks are a good uncertainty estimator in various applications (Although,

they are criticized for not being Bayesian). We show Ensembles of deterministic and stochastic Neural Networks can indeed be cast as an approximate Bayesian inference. Deep Ensembles have another weakness of having high space complexity, we provide an alternative to it by modifying the original Bayes by Backprop (BBB) algorithm to learn more general concrete mixture distributions over weights. We show our methods and its variants can give better uncertainty estimates at a significantly lower parametric overhead than Deep Ensembles. We validate our hypothesis through experiments like non-linear regression, predictive uncertainty estimation, detecting adversarial images and exploration-exploitation trade-off in reinforcement learning.

****

Best feature performance in codeswitched hate speech texts

Edward Ombui,Lawrence Muchemi,Peter Wagacha

How well can hate speech concept be abstracted in order to inform automatic classification in codeswitched texts by machine learning classifiers? We explore different representations and empirically evaluate their predictiveness using both conventional and deep learning algorithms in identifying hate speech in a ~48k human-annotated dataset that contain mixed languages, a phenomenon common among multilingual speakers. This paper espouses a novel approach to handle this challenge by introducing a hierarchical approach that employs Latent Dirichlet Allocation to generate topic models that feed into another high-level feature set that we acronym PDC. PDC groups similar meaning words in word families during the preprocessing stage for supervised learning models. The high-level PDC features generated are based on Ombui et al, (2019) hate speech annotation framework that is informed by the triangular theory of hate (Stanberg,2003). Results obtained from frequency-based models using the PDC feature on the annotated dataset of ~48k short messages comprising of tweets generated during the 2012 and 2017 Kenyan presidential elections indicate an improvement on classification accuracy in identifying hate speech as compared to the baseline

****

Geom-GCN: Geometric Graph Convolutional Networks

Hongbin Pei,Bingzhe Wei,Kevin Chen-Chuan Chang,Yu Lei,Bo Yang

Message-passing neural networks (MPNNs) have been successfully applied in a wide variety of applications in the real world. However, two fundamental weaknesses of MPNNs' aggregators limit their ability to represent graph-structured data: losing the structural information of nodes in neighborhoods and lacking the ability to capture long-range dependencies in disassortative graphs. Few studies have noticed the weaknesses from different perspectives. From the observations on classical neural network and network geometry, we propose a novel geometric aggregation scheme for graph neural networks to overcome the two weaknesses. The behind basic idea is the aggregation on a graph can benefit from a continuous space underlying the graph. The proposed aggregation scheme is permutation-invariant and consists of three modules, node embedding, structural neighborhood, and bi-level aggregation. We also present an implementation of the scheme in graph convolutional networks, termed Geom-GCN, to perform transductive learning on graphs. Experimental results show the proposed Geom-GCN achieved state-of-the-art performance on a wide range of open datasets of graphs.

****

Smart Ternary Quantization

Gregoire Morin,Ryan Razani,Vahid Partovi Nia,Eyyub Sari

Neural network models are resource hungry. Low bit quantization such as binary and ternary quantization is a common approach to alleviate this resource requirements. Ternary quantization provides a more flexible model and often beats binary quantization in terms of accuracy, but doubles memory and increases computation cost. Mixed quantization depth models, on another hand, allows a trade-off between accuracy and memory footprint. In such models, quantization depth is often chosen manually (which is a tiring task), or is tuned using a separate optimization routine (which requires training a quantized network multiple times). Here, w

e propose Smart Ternary Quantization (STQ) in which we modify the quantization d
epth directly through an adaptive regularization function, so that we train a mo
del only once. This method jumps between binary and ternary quantization while t
raining. We show its application on image classification.
****
HIPPOCAMPAL NEURONAL REPRESENTATIONS IN CONTINUAL LEARNING
Samia Mohinta,Rui Ponte Costa,Stephane Ciocchi
The hippocampus has long been associated with spatial memory and goal-directed
spatial navigation. However, the region’s independent role in continual learning
 of
navigational strategies has seldom been investigated. Here we analyse population
level
activity of hippocampal CA1 neurons in the context of continual learning of
two different spatial navigation strategies. Demixed Principal Component Analysi
s
(dPCA) is applied on neuronal recordings from 612 hippocampal CA1 neurons
of rodents learning to perform allocentric and egocentric spatial tasks. The com
ponents
uncovered using dPCA from the firing activity reveal that hippocampal
neurons encode relevant task variables such decisions, navigational strategies a
nd
reward location. We compare this hippocampal features with standard reinforcemen
t
learning algorithms, highlighting similarities and differences. Finally, we
demonstrate that a standard deep reinforcement learning model achieves similar
average performance when compared to animal learning, but fails to mimic animals
during task switching. Overall, our results gives insights into how the hippocam
pus
solves reinforced spatial continual learning, and puts forward a framework
to explicitly compare biological and machine learning during spatial continual
learning.
****
A GOODNESS OF FIT MEASURE FOR GENERATIVE NETWORKS
Lorenzo Luzi,Randall Balestriero,Richard Baraniuk
We define a goodness of fit measure for generative networks which captures how w
ell the network can generate the training data, which is necessary to learn the
true data distribution.
We demonstrate how our measure can be leveraged to understand mode collapse in g
enerative adversarial networks and provide practitioners with a novel way to per
form model comparison and early stopping without having to access another traine
d model as with Frechet Inception Distance or Inception Score. This measure show
s that several successful, popular generative models, such as DCGAN and WGAN, fa
ll very short of learning the data distribution. We identify this issue in gener
ative models and empirically show that overparameterization via subsampling data
 and using a mixture of models improves performance in terms of goodness of fit.
****
Gradients as Features for Deep Representation Learning
Fangzhou Mu,Yingyu Liang,Yin Li
We address the challenging problem of deep representation learning -- the effici
ent adaption of a pre-trained deep network to different tasks. Specifically, we
propose to explore gradient-based features. These features are gradients of the
model parameters with respect to a task-specific loss given an input sample. Our
 key innovation is the design of a linear model that incorporates both gradient
and activation of the pre-trained network. We demonstrate that our model provide
s a local linear approximation to an underlying deep model, and discuss importan
t theoretical insights. Moreover, we present an efficient algorithm for the trai
ning and inference of our model without computing the actual gradients. Our meth
od is evaluated across a number of representation-learning tasks on several data
sets and using different network architectures. Strong results are obtained in a
ll settings, and are well-aligned with our theoretical insights.

****
Deceptive Opponent Modeling with Proactive Network Interdiction for Stochastic Goal Recognition Control
Junren Luo,Wei Gao,Zhiyong Liao,Weilin Yuan,Wanpeng Zhang,Shaofei Chen
Goal recognition based on the observations of the behaviors collected online has been used to model some potential applications. Newly formulated problem of goal recognition design aims at facilitating the online goal recognition process by performing offline redesign of the underlying environment with hard action removal.
In this paper, we propose the stochastic goal recognition control (S-GRC) problem with two main stages: (1) deceptive opponent modeling based on maximum entropy regularized Markov decision processes (MDPs) and (2) goal recognition control under proactively static interdiction.
For the purpose of evaluation, we propose to use the worst case distinctiveness (wcd) as a measure of the non-distinctive path without revealing the true goals, the task of S-GRC is to interdict a set of actions that improve or reduce the wcd.
We empirically demonstrate that our proposed approach control the goal recognition process based on opponent's deceptive behavior.
****
Monotonic Multihead Attention
Xutai Ma,Juan Miguel Pino,James Cross,Liezl Puzon,Jiatao Gu
Simultaneous machine translation models start generating a target sequence before they have encoded or read the source sequence. Recent approach for this task either apply a fixed policy on transformer, or a learnable monotonic attention on a weaker recurrent neural network based structure. In this paper, we propose a new attention mechanism, Monotonic Multihead Attention (MMA), which introduced the monotonic attention mechanism to multihead attention. We also introduced two novel interpretable approaches for latency control that are specifically designed for multiple attentions. We apply MMA to the simultaneous machine translation task and demonstrate better latency-quality tradeoffs compared to MILk, the previous state-of-the-art approach.

****
Massively Multilingual Sparse Word Representations
Gábor Berend
In this paper, we introduce Mamus for constructing multilingual sparse word representations. Our algorithm operates by determining a shared set of semantic units which get reutilized across languages, providing it a competitive edge both in terms of speed and evaluation performance. We demonstrate that our proposed algorithm behaves competitively to strong baselines through a series of rigorous experiments performed towards downstream applications spanning over dependency parsing, document classification and natural language inference. Additionally, our experiments relying on the QVEC-CCA evaluation score suggests that the proposed sparse word representations convey an increased interpretability as opposed to alternative approaches. Finally, we are releasing our multilingual sparse word representations for the 27 typologically diverse set of languages that we conducted our various experiments on.
****
Attention over Phrases
Wanyun Cui
How to represent the sentence ``That's the last straw for her''? The answer of the self-attention is a weighted sum of each individual words, i.e. $$semantics=\alpha_1Emb(\text{That})+\alpha_2Emb(\text{'s})+\cdots+\alpha_nEmb(\text{her})$$. But the weighted sum of ``That's'', ``the'', ``last'', ``straw'' can hardly represent the semantics of the phrase. We argue that the phrases play an important role in attention.
If we combine some words into phrases, a more reasonable representation with compositions is
$$semantics=\alpha_1Emb(\text{That's})+Emb_2(\text{the last straw})+\alpha_3Emb(

\text{for})+\alpha_4Emb(\text{her})$$.
While recent studies prefer to use the attention mechanism to represent the natu ral language, few noticed the word compositions. In this paper, we study the pro blem of representing such compositional attentions in phrases. In this paper, we proposed a new attention architecture called HyperTransformer. Besides represen ting the words of the sentence, we introduce hypernodes to represent the candida te phrases in attention.
HyperTransformer has two phases. The first phase is used to attend over all word /phrase pairs, which is similar to the standard Transformer. The second phase is used to represent the inductive bias within each phrase. Specially, we incorpor ate the non-linear attention in the second phase. The non-linearity represents t he the semantic mutations in phrases. The experimental performance has been grea tly improved. In WMT16 English-German translation task, the BLEU increases from 20.90 (by Transformer) to 34.61 (by HyperTransformer).
****

Query-efficient Meta Attack to Deep Neural Networks
Jiawei Du,Hu Zhang,Joey Tianyi Zhou,Yi Yang,Jiashi Feng
Black-box attack methods aim to infer suitable attack patterns to targeted DNN m odels by only using output feedback of the models and the corresponding input qu eries. However, due to lack of prior and inefficiency in leveraging the query an d feedback information, existing methods are mostly query-intensive for obtainin g effective attack patterns. In this work, we propose a meta attack approach tha t is capable of attacking a targeted model with much fewer queries. Its high qu ery-efficiency stems from effective utilization of meta learning approaches in learning generalizable prior abstraction from the previously observed attack pat terns and exploiting such prior to help infer attack patterns from only a few q ueries and outputs. Extensive experiments on MNIST, CIFAR10 and tiny-Imagenet de monstrate that our meta-attack method can remarkably reduce the number of model queries without sacrificing the attack performance. Besides, the obtained meta a ttacker is not restricted to a particular model but can be used easily with a fa st adaptive ability to attack a variety of models. Our code will be released to the public.
****

BREAKING CERTIFIED DEFENSES: SEMANTIC ADVERSARIAL EXAMPLES WITH SPOOFED ROBUSTNESS CERTIFICATES
Amin Ghiasi,Ali Shafahi,Tom Goldstein
Defenses against adversarial attacks can be classified into certified and non-ce rtified. Certifiable defenses make networks robust within a certain ℓ_p-bou nded radius, so that it is impossible for the adversary to make adversarial exam ples in the certificate bound. We present an attack that maintains the impercept ibility property of adversarial examples while being outside of the certified ra dius. Furthermore, the proposed "Shadow Attack" can fool certifiably robust netw orks by producing an imperceptible adversarial example that gets misclassified a nd produces a strong ``spoofed'' certificate.
****

Meta-Learning Initializations for Image Segmentation
Sean M. Hendryx,Andrew B. Leach,Paul D. Hein,Clayton T. Morrison
While meta-learning approaches that utilize neural network representations have made progress in few-shot image classification, reinforcement learning, and, mor e recently, image semantic segmentation, the training algorithms and model archi tectures have become increasingly specialized to the few-shot domain. A natural question that arises is how to develop learning systems that scale from few-shot to many-shot settings while yielding human level performance in both. One scala ble potential approach that does not require ensembling many models nor the comp utational costs of relation networks, is to meta-learn an initialization. In thi s work, we study first-order meta-learning of initializations for deep neural ne tworks that must produce dense, structured predictions given an arbitrary amount of train- ing data for a new task. Our primary contributions include (1), an ex tension and experimental analysis of first-order model agnostic meta-learning al gorithms (including FOMAML and Reptile) to image segmentation, (2) a formalizati

on of the generalization error of episodic meta-learning algorithms, which we le verage to decrease error on unseen tasks, (3) a novel neural network architectur e built for parameter efficiency which we call EfficientLab, and (4) an empirica l study of how meta-learned initializations compare to ImageNet initializations as the training set size increases. We show that meta-learned initializations fo r image segmentation smoothly transition from canonical few-shot learning proble ms to larger datasets, outperforming random and ImageNet-trained initializations . Finally, we show both theoretically and empirically that a key limitation of M AML-type algorithms is that when adapting to new tasks, a single update procedur e is used that is not conditioned on the data. We find that our network, with an empirically estimated optimal update procedure yields state of the art results on the FSS-1000 dataset, while only requiring one forward pass through a single model at evaluation time.

****

Privacy-preserving Representation Learning by Disentanglement

Tassilo Klein,Moin Nabi

Deep learning and latest machine learning technology heralded an era of success in data analysis. Accompanied by the ever increasing performance, reaching super -human performance in many areas, is the requirement of amassing more and more d ata to train these models. Often ignored or underestimated, the big data curatio n is associated with the risk of privacy leakages. The proposed approach seeks t o mitigate these privacy issues. In order to sanitize data from sensitive conten t, we propose to learn a privacy-preserving data representation by disentangling into public and private part, with the public part being shareable without priv acy infringement. The proposed approach deals with the setting where the private features are not explicit, and is estimated though the course of learning. This is particularly appealing, when the notion of sensitive attribute is ``fuzzy''. We showcase feasibility in terms of classification of facial attributes and ide ntity on the CelebA dataset. The results suggest that private component can be r emoved in the cases where the the downstream task is known a priori (i.e., ``sup ervised''), and the case where it is not known a priori (i.e., ``weakly-supervis ed'').

****

An Exponential Learning Rate Schedule for Deep Learning

Zhiyuan Li,Sanjeev Arora

Intriguing empirical evidence exists that deep learning can work well with exoti c schedules for varying the learning rate. This paper suggests that the phenomen on may be due to Batch Normalization or BN(Ioffe & Szegedy, 2015), which is ubiq - uitous and provides benefits in optimization and generalization across all sta ndard architectures. The following new results are shown about BN with weight de cay and momentum (in other words, the typical use case which was not considered in earlier theoretical analyses of stand-alone BN (Ioffe & Szegedy, 2015; Santur kar et al., 2018; Arora et al., 2018)

• Training can be done using SGD with momentum and an exponentially in- creasing learning rate schedule, i.e., learning rate increases by some (1 + α) factor in every epoch for some $\alpha > 0$. (Precise statement in the paper.) To the best of ou r knowledge this is the first time such a rate schedule has been successfully us ed, let alone for highly successful architectures. As ex- pected, such training rapidly blows up network weights, but the net stays well-behaved due to normaliz ation.

• Mathematical explanation of the success of the above rate schedule: a rigor- o us proof that it is equivalent to the standard setting of BN + SGD + Standard Ra te Tuning + Weight Decay + Momentum. This equivalence holds for other normalizat ion layers as well, Group Normalization(Wu & He, 2018), Layer Normalization(Ba e t al., 2016), Instance Norm(Ulyanov et al., 2016), etc.

• A worked-out toy example illustrating the above linkage of hyper- parameters. Using either weight decay or BN alone reaches global minimum, but convergence fa ils when both are used.

****

End-to-end learning of energy-based representations for irregularly-sampled sign

als and images
Ronan Fablet,Lucas Drumetz,François Rousseau
For numerous domains, including for instance earth observation, medical imaging, astrophysics,..., available image and signal datasets often irregular space-time sampling patterns and large missing data rates. These sampling properties is a critical issue to apply state-of-the-art learning-based (e.g., auto-encoders, CNNs,...) to fully benefit from the available large-scale observations and reach breakthroughs in the reconstruction and identification of processes of interest. In this paper, we address the end-to-end learning of representations of signals, images and image sequences from irregularly-sampled data, {\em i.e.} when the training data involved missing data. From an analogy to Bayesian formulation, we consider energy-based representations. Two energy forms are investigated: one derived from auto-encoders and one relating to Gibbs energies. The learning stage of these energy-based representations (or priors) involve a joint interpolation issue, which resorts to solving an energy minimization problem under observation constraints. Using a neural-network-based implementation of the considered energy forms, we can state an end-to-end learning scheme from irregularly-sampled data. We demonstrate the relevance of the proposed representations for different case-studies: namely, multivariate time series, 2{\sc } images and image sequences.
****
Enabling Deep Spiking Neural Networks with Hybrid Conversion and Spike Timing Dependent Backpropagation
Nitin Rathi,Gopalakrishnan Srinivasan,Priyadarshini Panda,Kaushik Roy
Spiking Neural Networks (SNNs) operate with asynchronous discrete events (or spikes) which can potentially lead to higher energy-efficiency in neuromorphic hardware implementations. Many works have shown that an SNN for inference can be formed by copying the weights from a trained Artificial Neural Network (ANN) and setting the firing threshold for each layer as the maximum input received in that layer. These type of converted SNNs require a large number of time steps to achieve competitive accuracy which diminishes the energy savings. The number of time steps can be reduced by training SNNs with spike-based backpropagation from scratch, but that is computationally expensive and slow. To address these challenges, we present a computationally-efficient training technique for deep SNNs. We propose a hybrid training methodology: 1) take a converted SNN and use its weights and thresholds as an initialization step for spike-based backpropagation, and 2) perform incremental spike-timing dependent backpropagation (STDB) on this carefully initialized network to obtain an SNN that converges within few epochs and requires fewer time steps for input processing. STDB is performed with a novel surrogate gradient function defined using neuron's spike time. The weight update is proportional to the difference in spike timing between the current time step and the most recent time step the neuron generated an output spike. The SNNs trained with our hybrid conversion-and-STDB training perform at $10{\times}{-}25{\times}$ fewer number of time steps and achieve similar accuracy compared to purely converted SNNs. The proposed training methodology converges in less than 20 epochs of spike-based backpropagation for most standard image classification datasets, thereby greatly reducing the training complexity compared to training SNNs from scratch. We perform experiments on CIFAR-10, CIFAR-100 and ImageNet datasets for both VGG and ResNet architectures. We achieve top-1 accuracy of 65.19% for ImageNet dataset on SNN with 250 time steps, which is $10{\times}$ faster compared to converted SNNs with similar accuracy.
****
How to 0wn the NAS in Your Spare Time
Sanghyun Hong,Michael Davinroy,Yi■itcan Kaya,Dana Dachman-Soled,Tudor Dumitra■
New data processing pipelines and novel network architectures increasingly drive the success of deep learning. In consequence, the industry considers top-performing architectures as intellectual property and devotes considerable computational resources to discovering such architectures through neural architecture search (NAS). This provides an incentive for adversaries to steal these novel architectures; when used in the cloud, to provide Machine Learning as a Service (MLaaS)

, the adversaries also have an opportunity to reconstruct the architectures by e xploiting a range of hardware side-channels. However, it is challenging to recon struct novel architectures and pipelines without knowing the computational graph (e.g., the layers, branches or skip connections), the architectural parameters (e.g., the number of filters in a convolutional layer) or the specific pre-proce ssing steps (e.g. embeddings). In this paper, we design an algorithm that recons tructs the key components of a novel deep learning system by exploiting a small amount of information leakage from a cache side-channel attack, Flush+Reload. We use Flush+Reload to infer the trace of computations and the timing for each com putation. Our algorithm then generates candidate computational graphs from the t race and eliminates incompatible candidates through a parameter estimation proce ss. We implement our algorithm in PyTorch and Tensorflow. We demonstrate experim entally that we can reconstruct MalConv, a novel data pre-processing pipeline fo r malware detection, and ProxylessNAS-CPU, a novel network architecture for the ImageNet classification optimized to run on CPUs, without knowing the architectu re family. In both cases, we achieve 0% error. These results suggest hardware si de channels are a practical attack vector against MLaaS, and more efforts should be devoted to understanding their impact on the security of deep learning syste ms.

****

Generalized Zero-shot ICD Coding

Congzheng Song,Shanghang Zhang,Najmeh Sadoughi,Pengtao Xie,Eric Xing

The International Classification of Diseases (ICD) is a list of classification c odes for the diagnoses. Automatic ICD coding is in high demand as the manual cod ing can be labor-intensive and error-prone. It is a multi-label text classificat ion task with extremely long-tailed label distribution, making it difficult to p erform fine-grained classification on both frequent and zero-shot codes at the s ame time. In this paper, we propose a latent feature generation framework for ge neralized zero-shot ICD coding, where we aim to improve the prediction on codes that have no labeled data without compromising the performance on seen codes. Ou r framework generates pseudo features conditioned on the ICD code descriptions a nd exploits the ICD code hierarchical structure. To guarantee the semantic consi stency between the generated features and real features, we reconstruct the keyw ords in the input documents that are related to the conditioned ICD codes. To th e best of our knowledge, this works represents the first one that proposes an ad versarial generative model for the generalized zero-shot learning on multi-label text classification. Extensive experiments demonstrate the effectiveness of our approach. On the public MIMIC-III dataset, our methods improve the F1 score fro m nearly 0 to 20.91% for the zero-shot codes, and increase the AUC score by 3% (absolute improvement) from previous state of the art. We also show that the fram ework improves the performance on few-shot codes.

****

EXACT ANALYSIS OF CURVATURE CORRECTED LEARNING DYNAMICS IN DEEP LINEAR NETWORKS

Dongsung Huh

Deep neural networks exhibit complex learning dynamics due to the highly non-con vex loss landscape, which causes slow convergence and vanishing gradient problem s. Second order approaches, such as natural gradient descent, mitigate such prob lems by neutralizing the effect of potentially ill-conditioned curvature on the gradient-based updates, yet precise theoretical understanding on how such curvat ure correction affects the learning dynamics of deep networks has been lack- ing . Here, we analyze the dynamics of training deep neural networks under a general ized family of natural gradient methods that applies curvature corrections, and derive precise analytical solutions. Our analysis reveals that curvature correct ed update rules preserve many features of gradient descent, such that the learni ng trajectory of each singular mode in natural gradient descent follows precisel y the same path as gradient descent, while only accelerating the temporal dynami cs along the path. We also show that layer-restricted approximations of natural gradient, which are widely used in most second order methods (e.g. K-FAC), can s ignificantly distort the learning trajectory into highly diverging dynamics that significantly differs from true natural gradient, which may lead to undesirable

net- work properties. We also introduce fractional natural gradient that applies partial curvature correction, and show that it provides most of the benefit of full curvature correction in terms of convergence speed, with additional benefit of superior numerical stability and neutralizing vanishing/exploding gradient problems, which holds true also in layer-restricted approximations.
****
Learning to Reason: Distilling Hierarchy via Self-Supervision and Reinforcement Learning
Jung-Su Ha,Young-Jin Park,Hyeok-Joo Chae,Soon-Seo Park,Han-Lim Choi
We present a hierarchical planning and control framework that enables an agent to perform various tasks and adapt to a new task flexibly. Rather than learning an individual policy for each particular task, the proposed framework, DISH, distills a hierarchical policy from a set of tasks by self-supervision and reinforcement learning. The framework is based on the idea of latent variable models that represent high-dimensional observations using low-dimensional latent variables. The resulting policy consists of two levels of hierarchy: (i) a planning module that reasons a sequence of latent intentions that would lead to optimistic future and (ii) a feedback control policy, shared across the tasks, that executes the inferred intention. Because the reasoning is performed in low-dimensional latent space, the learned policy can immediately be used to solve or adapt to new tasks without additional training. We demonstrate the proposed framework can learn compact representations (3-dimensional latent states for a 90-dimensional humanoid system) while solving a small number of imitation tasks, and the resulting policy is directly applicable to other types of tasks, i.e., navigation in cluttered environments.
****
The Shape of Data: Intrinsic Distance for Data Distributions
Anton Tsitsulin,Marina Munkhoeva,Davide Mottin,Panagiotis Karras,Alex Bronstein,Ivan Oseledets,Emmanuel Mueller
The ability to represent and compare machine learning models is crucial in order to quantify subtle model changes, evaluate generative models, and gather insights on neural network architectures. Existing techniques for comparing data distributions focus on global data properties such as mean and covariance; in that sense, they are extrinsic and uni-scale. We develop a first-of-its-kind intrinsic and multi-scale method for characterizing and comparing data manifolds, using a lower-bound of the spectral variant of the Gromov-Wasserstein inter-manifold distance, which compares all data moments. In a thorough experimental study, we demonstrate that our method effectively discerns the structure of data manifolds even on unaligned data of different dimensionalities; moreover, we showcase its efficacy in evaluating the quality of generative models.
****
Measuring Numerical Common Sense: Is A Word Embedding Approach Effective?
Hiroaki Yamane,Chin-Yew Lin,Tatsuya Harada
Numerical common sense (e.g., ``a person with a height of 2m is very tall'') is essential when deploying artificial intelligence (AI) systems in society. To predict ranges of small and large values for a given target noun and unit, previous studies have implemented a rule-based method that processed numeric values appearing in a natural language by using template matching. To obtain numerical knowledge, crawled textual data from web pages are frequently used as the input in the above method. Although this is an important task, few studies have addressed the availability of numerical common sense extracted from corresponding textual information. To this end, we first used a crowdsourcing service to obtain sufficient data for a subjective agreement on numerical common sense. Second, to examine whether common sense is attributed to current word embedding, we examined the performance of a regressor trained on the obtained data. In comparison with humans, the performance of an automatic relevance determination regression model was good, particularly when the unit was yen (a maximum correlation coefficient of 0.57). Although all the regression approach with word embedding does not predict values with high correlation coefficients, this word-embedding method could potentially contribute to construct numerical common sense for AI deployme

nt.
****
Learning DNA folding patterns with Recurrent Neural Networks
Michal Rozenwald,Aleksandra Galitsyna,Ekaterina Khrameeva,Grigory Sapunov,Mikhail S. Gelfand

The recent expansion of machine learning applications to molecular biology proved to have a significant contribution to our understanding of biological systems, and genome functioning in particular. Technological advances enabled the collection of large epigenetic datasets, including information about various DNA binding factors (ChIP-Seq) and DNA spatial structure (Hi-C). Several studies have confirmed the correlation between DNA binding factors and Topologically Associating Domains (TADs) in DNA structure. However, the information about physical proximity represented by genomic coordinate was not yet used for the improvement of the prediction models.

In this research, we focus on Machine Learning methods for prediction of folding patterns of DNA in a classical model organism Drosophila melanogaster. The paper considers linear models with four types of regularization, Gradient Boosting and Recurrent Neural Networks for the prediction of chromatin folding patterns from epigenetic marks. The bidirectional LSTM RNN model outperformed all the models and gained the best prediction scores. This demonstrates the utilization of complex models and the importance of memory of sequential DNA states for the chromatin folding. We identify informative epigenetic features that lead to the further conclusion of their biological significance.
****
Generative Adversarial Nets for Multiple Text Corpora
Diego Klabjan,Baiyang Wang
Generative adversarial nets (GANs) have been successfully applied to the artificial generation of image data. In terms of text data, much has been done on the artificial generation of natural language from a single corpus. We consider multiple text corpora as the input data, for which there can be two applications of GANs: (1) the creation of consistent cross-corpus word embeddings given different word embeddings per corpus; (2) the generation of robust bag-of-words document embeddings for each corpora. We demonstrate our GAN models on real-world text data sets from different corpora, and show that embeddings from both models lead to improvements in supervised learning problems.
****
Understanding Generalization in Recurrent Neural Networks
Zhuozhuo Tu,Fengxiang He,Dacheng Tao
In this work, we develop the theory for analyzing the generalization performance of recurrent neural networks. We first present a new generalization bound for recurrent neural networks based on matrix 1-norm and Fisher-Rao norm. The definition of Fisher-Rao norm relies on a structural lemma about the gradient of RNNs. This new generalization bound assumes that the covariance matrix of the input data is positive definite, which might limit its use in practice. To address this issue, we propose to add random noise to the input data and prove a generalization bound for training with random noise, which is an extension of the former one. Compared with existing results, our generalization bounds have no explicit dependency on the size of networks. We also discover that Fisher-Rao norm for RNNs can be interpreted as a measure of gradient, and incorporating this gradient measure not only can tighten the bound, but allows us to build a relationship between generalization and trainability. Based on the bound, we theoretically analyze the effect of covariance of features on generalization of RNNs and discuss how weight decay and gradient clipping in the training can help improve generalization.
****
Weakly-Supervised Trajectory Segmentation for Learning Reusable Skills
Parsa Mahmoudieh,Trevor Darrell,Deepak Pathak
Learning useful and reusable skill, or sub-task primitives, is a long-standing p

roblem in sensorimotor control. This is challenging because it's hard to define what constitutes a useful skill. Instead of direct manual supervision which is tedious and prone to bias, in this work, our goal is to extract reusable skills from a collection of human demonstrations collected directly for several end-tasks. We propose a weakly-supervised approach for trajectory segmentation following the classic work on multiple instance learning. Our approach is end-to-end trainable, works directly from high-dimensional input (e.g., images) and only requires the knowledge of what skill primitives are present at training, without any need of segmentation or ordering of primitives. We evaluate our approach via rigorous experimentation across four environments ranging from simulation to real world robots, procedurally generated to human collected demonstrations and discrete to continuous action space. Finally, we leverage the generated skill segmentation to demonstrate preliminary evidence of zero-shot transfer to new combinations of skills. Result videos at https://sites.google.com/view/trajectory-segmentation/

****

Learn Interpretable Word Embeddings Efficiently with von Mises-Fisher Distribution

Minghong Yao,Liansheng Zhuang,Houqiang Li,Jian Yang,Shafei Wang

Word embedding plays a key role in various tasks of natural language processing. However, the dominant word embedding models don't explain what information is carried with the resulting embeddings. To generate interpretable word embeddings we intend to replace the word vector with a probability density distribution. The insight here is that if we regularize the mixture distribution of all words to be uniform, then we can prove that the inner product between word embeddings represent the point-wise mutual information between words. Moreover, our model can also handle polysemy. Each word's probability density distribution will generate different vectors for its various meanings. We have evaluated our model in several word similarity tasks. Results show that our model can outperform the dominant models consistently in these tasks.

****

Goten: GPU-Outsourcing Trusted Execution of Neural Network Training and Prediction

Lucien K.L. Ng,Sherman S.M. Chow,Anna P.Y. Woo,Donald P. H. Wong,Yongjun Zhao

Before we saw worldwide collaborative efforts in training machine-learning models or widespread deployments of prediction-as-a-service, we need to devise an ef■cient privacy-preserving mechanism which guarantees the privacy of all stakeholders (data contributors, model owner, and queriers). Slaom (ICLR ’19) preserves privacy only for prediction by leveraging both trusted environment (e.g., Intel SGX) and untrusted GPU. The challenges for enabling private training are explicitly left open – its pre-computation technique does not hide the model weights and fails to support dynamic quantization corresponding to the large changes in weight magnitudes during training. Moreover, it is not a truly outsourcing solution since (of■ine) pre-computation for a job takes as much time as computing the job locally by SGX, i.e., it only works before all pre-computations are exhausted.

We propose Goten, a privacy-preserving framework supporting both training and prediction. We tackle all the above challenges by proposing a secure outsourcing protocol which 1) supports dynamic quantization, 2) hides the model weight from GPU, and 3) performs better than a pure-SGX solution even if we perform the precomputation online. Our solution leverages a non-colluding assumption which is often employed by cryptographic solutions aiming for practical ef■ciency (IEEE SP ’13, Usenix Security ’17, PoPETs ’19). We use three servers, which can be reduced to two if the pre-computation is done of■ine. Furthermore, we implement our tailor-made memory-aware measures for minimizing the overhead when the SGX memory limit is exceeded (cf., EuroSys ’17, Usenix ATC ’19). Compared to a pure-SGX solution, our experiments show that Goten can speed up linear-layer computations in VGG up to 40×, and overall speed up by 8.64× on VGG11.

****

Limitations for Learning from Point Clouds

Christian Bueno,Alan G. Hylton
In this paper we prove new universal approximation theorems for deep learning on point clouds that do not assume fixed cardinality. We do this by first generalizing the classical universal approximation theorem to general compact Hausdorff spaces and then applying this to the permutation-invariant architectures presented in 'PointNet' (Qi et al) and 'Deep Sets' (Zaheer et al). Moreover, though both architectures operate on the same domain, we show that the constant functions are the only functions they can mutually uniformly approximate. In particular, DeepSets architectures cannot uniformly approximate the diameter function but can uniformly approximate the center of mass function but it is the other way around for PointNet.
****

Conservative Uncertainty Estimation By Fitting Prior Networks
Kamil Ciosek,Vincent Fortuin,Ryota Tomioka,Katja Hofmann,Richard Turner
Obtaining high-quality uncertainty estimates is essential for many applications of deep neural networks. In this paper, we theoretically justify a scheme for estimating uncertainties, based on sampling from a prior distribution. Crucially, the uncertainty estimates are shown to be conservative in the sense that they never underestimate a posterior uncertainty obtained by a hypothetical Bayesian algorithm. We also show concentration, implying that the uncertainty estimates converge to zero as we get more data. Uncertainty estimates obtained from random priors can be adapted to any deep network architecture and trained using standard supervised learning pipelines. We provide experimental evaluation of random priors on calibration and out-of-distribution detection on typical computer vision tasks, demonstrating that they outperform deep ensembles in practice.
****

Re-Examining Linear Embeddings for High-dimensional Bayesian Optimization
Benjamin Letham,Roberto Calandra,Akshara Rai,Eytan Bakshy
Bayesian optimization (BO) is a popular approach to optimize resource-intensive black-box functions.
A significant challenge in BO is to scale to high-dimensional parameter spaces while retaining sample efficiency.
A solution considered in previous literature is to embed the high-dimensional parameter space into a lower-dimensional manifold, often a random linear embedding. In this paper, we identify several crucial issues and misconceptions about the use of linear embeddings for BO. We thoroughly study and analyze the consequences of using linear embeddings and show that some of the design choices in current approaches adversely impact their performance. Based on this new theoretical understanding we propose ALEBO, a new algorithm for high-dimensional BO via linear embeddings that outperforms state-of-the-art methods on a range of problems.
****

ASYNCHRONOUS MULTI-AGENT GENERATIVE ADVERSARIAL IMITATION LEARNING
Xin Zhang,Weixiao Huang,Renjie Liao,Yanhua Li
Imitation learning aims to inversely learn a policy from expert demonstrations, which has been extensively studied in the literature for both single-agent setting with Markov decision process (MDP) model, and multi-agent setting with Markov game (MG) model. However, existing approaches for general multi-agent Markov games are not applicable to multi-agent extensive Markov games, where agents make asynchronous decisions following a certain order, rather than simultaneous decisions. We propose a novel framework for asynchronous multi-agent generative adversarial imitation learning (AMAGAIL) under general extensive Markov game settings, and the learned expert policies are proven to guarantee subgame perfect equilibrium (SPE), a more general and stronger equilibrium than Nash equilibrium (NE). The experiment results demonstrate that compared to state-of-the-art baselines, our AMAGAIL model can better infer the policy of each expert agent using their demonstration data collected from asynchronous decision-making scenarios (i.e., extensive Markov games).
****

Predictive Coding for Boosting Deep Reinforcement Learning with Sparse Rewards
Xingyu Lu,Pieter Abbeel,Stas Tiomkin

While recent progress in deep reinforcement learning has enabled robots to learn complex behaviors, tasks with long horizons and sparse rewards remain an ongoing challenge. In this work, we propose an effective reward shaping method through predictive coding to tackle sparse reward problems. By learning predictive representations offline and using these representations for reward shaping, we gain access to reward signals that understand the structure and dynamics of the environment. In particular, our method achieves better learning by providing reward signals that 1) understand environment dynamics 2) emphasize on features most useful for learning 3) resist noise in learned representations through reward accumulation. We demonstrate the usefulness of this approach in different domains ranging from robotic manipulation to navigation, and we show that reward signals produced through predictive coding are as effective for learning as hand-crafted rewards.
****

NORML: Nodal Optimization for Recurrent Meta-Learning

David van Niekerk

Meta-learning is an exciting and powerful paradigm that aims to improve the effectiveness of current learning systems. By formulating the learning process as an optimization problem, a model can learn how to learn while requiring significantly less data or experience than traditional approaches. Gradient-based meta-learning methods aims to do just that, however recent work have shown that the effectiveness of these approaches are primarily due to feature reuse and very little has to do with priming the system for rapid learning (learning to make effective weight updates on unseen data distributions). This work introduces Nodal Optimization for Recurrent Meta-Learning (NORML), a novel meta-learning framework where an LSTM-based meta-learner performs neuron-wise optimization on a learner for efficient task learning. Crucially, the number of meta-learner parameters needed in NORML, increases linearly relative to the number of learner parameters. Allowing NORML to potentially scale to learner networks with very large numbers of parameters. While NORML also benefits from feature reuse it is shown experimentally that the meta-learner LSTM learns to make effective weight updates using information from previous data-points and update steps.
****

Keyword Spotter Model for Crop Pest and Disease Monitoring from Community Radio Data

Benjamin Akera,Joyce Nakatumba-Nabende,Ali Hussein,Daniel Ssendiwala,Jonathan Mukiibi

In societies with well developed internet infrastructure, social media is the leading medium of communication for various social issues especially for breaking news situations. In rural Uganda however, public community radio is still a dominant means for news dissemination. Community radio gives audience to the general public especially to individuals living in rural areas, and thus plays an important role in giving a voice to those living in the broadcast area. It is an avenue for participatory communication and a tool relevant in both economic and social development.This is supported by the rise to ubiquity of mobile phones providing access to phone-in or text-in talk shows. In this paper, we describe an approach to analysing the readily available community radio data with machine learning-based speech keyword spotting techniques. We identify the keywords of interest related to agriculture and build models to automatically identify these keywords from audio streams. Our contribution through these techniques is a cost-efficient and effective way to monitor food security concerns particularly in rural areas. Through keyword spotting and radio talk show analysis, issues such as crop diseases, pests, drought and famine can be captured and fed into an early warning system for stakeholders and policy makers.
****

NAS-Bench-1Shot1: Benchmarking and Dissecting One-shot Neural Architecture Search

Arber Zela,Julien Siems,Frank Hutter

One-shot neural architecture search (NAS) has played a crucial role in making NAS methods computationally feasible in practice. Nevertheless, there is still a

lack of understanding on how these weight-sharing algorithms exactly work due to the many factors controlling the dynamics of the process. In order to allow a scientific study of these components, we introduce a general framework for one-shot NAS that can be instantiated to many recently-introduced variants and introduce a general benchmarking framework that draws on the recent large-scale tabular benchmark NAS-Bench-101 for cheap anytime evaluations of one-shot NAS methods. To showcase the framework, we compare several state-of-the-art one-shot NAS methods, examine how sensitive they are to their hyperparameters and how they can be improved by tuning their hyperparameters, and compare their performance to that of blackbox optimizers for NAS-Bench-101.
****

Defense against Adversarial Examples by Encoder-Assisted Search in the Latent Coding Space
Wenjing Huang,Shikui Tu,Lei Xu
Deep neural networks were shown to be vulnerable to crafted adversarial perturbations, and thus bring serious safety problems. To solve this problem, we proposed $\text{AE-GAN}_\text{+sr}$, a framework for purifying input images by searching a closest natural reconstruction with little computation. We first build a reconstruction network AE-GAN, which adapted auto-encoder by introducing adversarial loss to the objective function. In this way, we can enhance the generative ability of decoder and preserve the abstraction ability of encoder to form a self-organized latent space. In the inference time, when given an input, we will start a search process in the latent space which aims to find the closest reconstruction to the given image on the distribution of normal data. The encoder can provide a good start point for the searching process, which saves much computation cost. Experiments show that our method is robust against various attacks and can reach comparable even better performance to similar methods with much fewer computations.
****

Conditional generation of molecules from disentangled representations
Amina Mollaysa,Brooks Paige,Alexandros Kalousis
Though machine learning approaches have shown great success in estimating properties of small molecules, the inverse problem of generating molecules with desired properties remains challenging. This difficulty is in part because the set of molecules which have a given property is structurally very diverse. Treating this inverse problem as a conditional distribution estimation task, we draw upon work in learning disentangled representations to learn a conditional distribution over molecules given a desired property, where the molecular structure is encoded in a continuous latent random variable. By including property information as an input factor independent from the structure representation, one can perform conditional molecule generation via a ``style transfer'' process, in which we explicitly set the property to a desired value at generation time. In contrast to existing approaches, we disentangle the latent factors from the property factors using a regularization term which constrains the generated molecules to have the property provided to the generation network, no matter how the latent factor changes.
****

Learning RNNs with Commutative State Transitions
Edo Cohen-Karlik,Amir Globerson
Many machine learning tasks involve analysis of set valued inputs, and thus the learned functions are expected to be permutation invariant. Recent works (e.g., Deep Sets) have sought to characterize the neural architectures which result in permutation invariance. These typically correspond to applying the same pointwise function to all set components, followed by sum aggregation. Here we take a different approach to such architectures and focus on recursive architectures such as RNNs, which are not permutation invariant in general, but can implement permutation invariant functions in a very compact manner. We first show that commutativity and associativity of the state transition function result in permutation invariance. Next, we derive a regularizer that minimizes the degree of non-commutativity in the transitions. Finally, we demonstrate tha

t the resulting method outperforms other methods for learning permutation invari
ant models, due to its use of recursive computation.
****

XD: Cross-lingual Knowledge Distillation for Polyglot Sentence Embeddings
Maksym Del,Mark Fishel
Current state-of-the-art results in multilingual natural language inference (NLI
) are based on tuning XLM (a pre-trained polyglot language model) separately for
 each language involved, resulting in multiple models. We reach significantly hi
gher NLI results with a single model for all languages via multilingual tuning.
Furthermore, we introduce cross-lingual knowledge distillation (XD), where the s
ame polyglot model is used both as teacher and student across languages to impro
ve its sentence representations without using the end-task labels. When used alo
ne, XD beats multilingual tuning for some languages and the combination of them
both results in a new state-of-the-art of 79.2% on the XNLI dataset, surpassing
the previous result by absolute 2.5%. The models and code for reproducing our ex
periments will be made publicly available after de-anonymization.
****

LAVAE: Disentangling Location and Appearance
Andrea Dittadi,Ole Winther
We propose a probabilistic generative model for unsupervised learning of structu
red, interpretable, object-based representations of visual scenes. We use amorti
zed variational inference to train the generative model end-to-end. The learned
representations of object location and appearance are fully disentangled, and ob
jects are represented independently of each other in the latent space. Unlike pr
evious approaches that disentangle location and appearance, ours generalizes sea
mlessly to scenes with many more objects than encountered in the training regime
. We evaluate the proposed model on multi-MNIST and multi-dSprites data sets.
****

Sparse Skill Coding: Learning Behavioral Hierarchies with Sparse Codes
Sophia Sanborn,Michael Chang,Sergey Levine,Thomas Griffiths
Many approaches to hierarchical reinforcement learning aim to identify sub-goal
structure in tasks. We consider an alternative perspective based on identifying
behavioral `motifs'---repeated action sequences that can be compressed to yield
a compact code of action trajectories. We present a method for iteratively compr
essing action trajectories to learn nested behavioral hierarchies of arbitrary d
epth, with actions of arbitrary length. The learned temporally extended actions
provide new action primitives that can participate in deeper hierarchies as the
agent learns. We demonstrate the relevance of this approach for tasks with non-t
rivial hierarchical structure and show that the approach can be used to accelera
te learning in recursively more complex tasks through transfer.
****

REFINING MONTE CARLO TREE SEARCH AGENTS BY MONTE CARLO TREE SEARCH
Katsuki Ohto
Reinforcement learning methods that continuously learn neural networks by episod
e generation with game tree search have been successful in two-person complete i
nformation deterministic games such as chess, shogi, and Go. However, there are
only reports of practical cases and there are little evidence to guarantee the s
tability and the final performance of learning process. In this research, the co
ordination of episode generation was focused on. By means of regarding the entir
e system as game tree search, the new method can handle the trade-off between ex
ploitation and exploration during episode generation. The experiments with a sma
ll problem showed that it had robust performance compared to the existing method
, Alpha Zero.
****

A Bilingual Generative Transformer for Semantic Sentence Embedding
John Wieting,Graham Neubig,Taylor Berg-Kirkpatrick
Semantic sentence embedding models take natural language sentences and turn them
 into vectors, such that similar vectors indicate similarity in the semantics be
tween the sentences. Bilingual data offers a useful signal for learning such emb
eddings: properties shared by both sentences in a translation pair are likely se

mantic, while divergent properties are likely stylistic or language-specific. We propose a deep latent variable model that attempts to perform source separation on parallel sentences, isolating what they have in common in a latent semantic vector, and explaining what is left over with language-specific latent vectors. Our proposed approach differs from past work on semantic sentence encoding in two ways. First, by using a variational probabilistic framework, we introduce priors that encourage source separation, and can use our model's posterior to predict sentence embeddings for monolingual data at test time. Second, we use high- capacity transformers as both data generating distributions and inference networks – contrasting with most past work on sentence embeddings. In experiments, our approach substantially outperforms the state-of-the-art on a standard suite of se- mantic similarity evaluations. Further, we demonstrate that our approach yields the largest gains on more difficult subsets of test where simple word overlap is not a good indicator of similarity.
****

Learning to Coordinate Manipulation Skills via Skill Behavior Diversification
Youngwoon Lee,Jingyun Yang,Joseph J. Lim
When mastering a complex manipulation task, humans often decompose the task into sub-skills of their body parts, practice the sub-skills independently, and then execute the sub-skills together. Similarly, a robot with multiple end-effectors can perform complex tasks by coordinating sub-skills of each end-effector. To realize temporal and behavioral coordination of skills, we propose a modular framework that first individually trains sub-skills of each end-effector with skill behavior diversification, and then learns to coordinate end-effectors using diverse behaviors of the skills. We demonstrate that our proposed framework is able to efficiently coordinate skills to solve challenging collaborative control tasks such as picking up a long bar, placing a block inside a container while pushing the container with two robot arms, and pushing a box with two ant agents. Videos and code are available at https://clvrai.com/coordination
****

DeepPCM: Predicting Protein-Ligand Binding using Unsupervised Learned Representations
Paul Kim,Robin Winter,Djork-Arné Clevert
In-silico protein-ligand binding prediction is an ongoing area of research in computational chemistry and machine learning based drug discovery, as an accurate predictive model could greatly reduce the time and resources necessary for the detection and prioritization of possible drug candidates. Proteochemometric modeling (PCM) attempts to make an accurate model of the protein-ligand interaction space by combining explicit protein and ligand descriptors. This requires the creation of information-rich, uniform and computer interpretable representations of proteins and ligands. Previous work in PCM modeling relies on pre-defined, hand crafted feature extraction methods, and many methods use protein descriptors that require alignment or are otherwise specific to a particular group of related proteins. However, recent advances in representation learning have shown that unsupervised machine learning can be used to generate embeddings which outperform complex, human-engineered representations. We apply this reasoning to propose a novel proteochemometric modeling methodology which, for the first time, uses embeddings generated via unsupervised representation learning for both the protein and ligand descriptors. We evaluate performance on various splits of a benchmark dataset, including a challenging split that tests the model's ability to generalize to proteins for which bioactivity data is greatly limited, and we find that our method consistently outperforms state-of-the-art methods.
****

Ternary MobileNets via Per-Layer Hybrid Filter Banks
Dibakar Gope,Jesse G Beu,Urmish Thakker,Matthew Mattina
MobileNets family of computer vision neural networks have fueled tremendous progress in the design and organization of resource-efficient architectures in recent years. New applications with stringent real-time requirements in highly constrained devices require further compression of MobileNets-like already computeefficient networks. Model quantization is a widely used technique to compress and ac

celerate neural network inference and prior works have quantized MobileNets to 4 − 6 bits albeit with a modest to significant drop in accuracy. While quantization to sub-byte values (i.e. precision ≤ 8 bits) has been valuable, even further quantization of MobileNets to binary or ternary values is necessary to realize significant energy savings and possibly runtime speedups on specialized hardware, such as ASICs and FPGAs. Under the key observation that convolutional filters at each layer of a deep neural network may respond differently to ternary quantization, we propose a novel quantization method that generates per-layer hybrid filter banks consisting of full-precision and ternary weight filters for MobileNets. The layer-wise hybrid filter banks essentially combine the strengths of full-precision and ternary weight filters to derive a compact, energy-efficient architecture for MobileNets. Using this proposed quantization method, we quantized a substantial portion of weight filters of MobileNets to ternary values resulting in 27.98% savings in energy, and a 51.07% reduction in the model size, while achieving comparable accuracy and no degradation in throughput on specialized hardware in comparison to the baseline full-precision MobileNets.
****

Constant Curvature Graph Convolutional Networks
Gregor Bachmann,Gary Bécigneul,Octavian-Eugen Ganea
Interest has been rising lately towards methods representing data in non-Euclidean spaces, e.g. hyperbolic or spherical. These geometries provide specific inductive biases useful for certain real-world data properties, e.g. scale-free or hierarchical graphs are best embedded in a hyperbolic space. However, the very popular class of graph neural networks is currently limited to model data only via Euclidean node embeddings and associated vector space operations. In this work, we bridge this gap by proposing mathematically grounded generalizations of graph convolutional networks (GCN) to (products of) constant curvature spaces. We do this by i) extending the gyro-vector space theory from hyperbolic to spherical spaces, providing a unified and smooth view of the two geometries, ii) leveraging gyro-barycentric coordinates that generalize the classic Euclidean concept of the center of mass. Our class of models gives strict generalizations in the sense that they recover their Euclidean counterparts when the curvature goes to zero from either side. Empirically, our methods outperform different types of classic Euclidean GCNs in the tasks of node classification and minimizing distortion for symbolic data exhibiting non-Euclidean behavior, according to their discrete curvature.
****

Variational Information Bottleneck for Unsupervised Clustering: Deep Gaussian Mixture Embedding
Yigit Ugur,George Arvanitakis,Abdellatif Zaidi
In this paper, we develop an unsupervised generative clustering framework that combines variational information bottleneck and the Gaussian Mixture Model. Specifically, in our approach we use the variational information bottleneck method and model the latent space as a mixture of Gaussians. We derive a bound on the cost function of our model that generalizes the evidence lower bound (ELBO); and provide a variational inference type algorithm that allows to compute it. In the algorithm, the coders' mappings are parametrized using neural networks and the bound is approximated by Markov sampling and optimized with stochastic gradient descent. Numerical results on real datasets are provided to support the efficiency of our method.
****

Combining graph and sequence information to learn protein representations
Hassan Kané,Mohamed Coulibali,Pelkins Ajanoh,Ali Abdalla
Computational methods that infer the function of proteins are key to understanding life at the molecular level. In recent years, representation learning has emerged as a powerful paradigm to discover new patterns among entities as varied as images, words, speech, molecules. In typical representation learning, there is only one source of data or one level of abstraction at which the learned representation occurs. However, proteins can be described by their primary, secondary, tertiary, and quaternary structure or even as nodes in protein-protein interacti

on networks. Given that protein function is an emergent property of all these le
vels of interactions in this work, we learn joint representations from both amin
o acid sequence and multilayer networks representing tissue-specific protein-pro
tein interactions. Using these representations, we train machine learning models
 that outperform existing methods on the task of tissue-specific protein functio
n prediction on 10 out of 13 tissues. Furthermore, we outperform existing method
s by 19% on average.
****
FINBERT: FINANCIAL SENTIMENT ANALYSIS WITH PRE-TRAINED LANGUAGE MODELS
Dogu Araci,Zulkuf Genc
While many sentiment classification solutions report high accuracy scores in pr
oduct or movie review datasets, the performance of the methods in niche domains
such as finance still largely falls behind. The reason of this gap is the domain
-specific language, which decreases the applicability of existing models, and la
ck of quality labeled data to learn the new context of positive and negative in
the specific domain. Transfer learning has been shown to be successful in adapti
ng to new domains without large training data sets. In this paper, we explore th
e effectiveness of NLP transfer learning in financial sentiment classification.
We introduce FinBERT, a language model based on BERT, which improved the state-o
f-the-art performance by 14 percentage points for a financial sentiment classifi
cation task in FinancialPhrasebank dataset.
****
Robust Subspace Recovery Layer for Unsupervised Anomaly Detection
Chieh-Hsin Lai,Dongmian Zou,Gilad Lerman
We propose a neural network for unsupervised anomaly detection with a novel robu
st subspace recovery layer (RSR layer). This layer seeks to extract the underlyi
ng subspace from a latent representation of the given data and removes outliers
that lie away from this subspace. It is used within an autoencoder. The encoder
maps the data into a latent space, from which the RSR layer extracts the subspac
e. The decoder then smoothly maps back the underlying subspace to a ``manifold"
close to the original inliers. Inliers and outliers are distinguished according
to the distances between the original and mapped positions (small for inliers an
d large for outliers). Extensive numerical experiments with both image and docum
ent datasets demonstrate state-of-the-art precision and recall.
****
Learning Nearly Decomposable Value Functions Via Communication Minimization
Tonghan Wang*,Jianhao Wang*,Chongyi Zheng,Chongjie Zhang
Reinforcement learning encounters major challenges in multi-agent settings, such
 as scalability and non-stationarity. Recently, value function factorization lea
rning emerges as a promising way to address these challenges in collaborative mu
lti-agent systems. However, existing methods have been focusing on learning full
y decentralized value functions, which are not efficient for tasks requiring com
munication. To address this limitation, this paper presents a novel framework fo
r learning nearly decomposable Q-functions (NDQ) via communication minimization,
 with which agents act on their own most of the time but occasionally send messa
ges to other agents in order for effective coordination. This framework hybridiz
es value function factorization learning and communication learning by introduci
ng two information-theoretic regularizers. These regularizers are maximizing mut
ual information between agents' action selection and communication messages whil
e minimizing the entropy of messages between agents. We show how to optimize the
se regularizers in a way that is easily integrated with existing value function
factorization methods such as QMIX. Finally, we demonstrate that, on the StarCra
ft unit micromanagement benchmark, our framework significantly outperforms basel
ine methods and allows us to cut off more than 80% of communication without s
acrificing the performance. The videos of our experiments are available at https
://sites.google.com/view/ndq.
****
Batch Normalization is a Cause of Adversarial Vulnerability
Angus Galloway,Anna Golubeva,Thomas Tanay,Medhat Moussa,Graham W. Taylor
Batch normalization (BN) is often used in an attempt to stabilize and accelerate

training in deep neural networks. In many cases it indeed decreases the number of parameter updates required to achieve low training error. However, it also reduces robustness to small adversarial input perturbations and common corruptions by double-digit percentages, as we show on five standard datasets. Furthermore, we find that substituting weight decay for BN is sufficient to nullify a relationship between adversarial vulnerability and the input dimension. A recent mean-field analysis found that BN induces gradient explosion when used on multiple layers, but this cannot fully explain the vulnerability we observe, given that it occurs already for a single BN layer. We argue that the actual cause is the tilting of the decision boundary with respect to the nearest-centroid classifier along input dimensions of low variance. As a result, the constant introduced for numerical stability in the BN step acts as an important hyperparameter that can be tuned to recover some robustness at the cost of standard test accuracy. We explain this mechanism explicitly on a linear ``toy model and show in experiments that it still holds for nonlinear ``real-world models.

****

Undersensitivity in Neural Reading Comprehension

Johannes Welbl,Pasquale Minervini,Max Bartolo,Pontus Stenetorp,Sebastian Riedel

Neural reading comprehension models have recently achieved impressive gener- alisation results, yet still perform poorly when given adversarially selected input. Most prior work has studied semantically invariant text perturbations which cause a model’s prediction to change when it should not. In this work we focus on the complementary problem: excessive prediction undersensitivity where input text is meaningfully changed, and the model’s prediction does not change when it should. We formulate a noisy adversarial attack which searches among semantic variations of comprehension questions for which a model still erroneously pro- duces the same answer as the original question – and with an even higher prob- ability. We show that – despite comprising unanswerable questions – SQuAD2.0 and NewsQA models are vulnerable to this attack and commit a substantial frac- tion of errors on adversarially generated questions. This indicates that current models—even where they can correctly predict the answer—rely on spurious sur- face patterns and are not necessarily aware of all information provided in a given comprehension question. Developing this further, we experiment with both data augmentation and adversarial training as defence strategies: both are able to sub- stantially decrease a model’s vulnerability to undersensitivity attacks on held out evaluation data. Finally, we demonstrate that adversarially robust models gener- alise better in a biased data setting with a train/evaluation distribution mismatch; they are less prone to overly rely on predictive cues only present in the training set and outperform a conventional model in the biased data setting by up to 11% F1.

****

Extreme Classification via Adversarial Softmax Approximation

Robert Bamler,Stephan Mandt

Training a classifier over a large number of classes, known as 'extreme classification', has become a topic of major interest with applications in technology, science, and e-commerce. Traditional softmax regression induces a gradient cost proportional to the number of classes C, which often is prohibitively expensive. A popular scalable softmax approximation relies on uniform negative sampling, which suffers from slow convergence due a poor signal-to-noise ratio. In this paper, we propose a simple training method for drastically enhancing the gradient signal by drawing negative samples from an adversarial model that mimics the data distribution. Our contributions are three-fold: (i) an adversarial sampling mechanism that produces negative samples at a cost only logarithmic in C, thus still resulting in cheap gradient updates; (ii) a mathematical proof that this adversarial sampling minimizes the gradient variance while any bias due to non-uniform sampling can be removed; (iii) experimental results on large scale data sets that show a reduction of the training time by an order of magnitude relative to several competitive baselines.

****

IS THE LABEL TRUSTFUL: TRAINING BETTER DEEP LEARNING MODEL VIA UNCERTAINTY MININ G NET
Yang Sun,Abhishek Kolagunda,Steven Eliuk,Xiaolong Wang
In this work, we consider a new problem of training deep neural network on parti ally labeled data with label noise. As far as we know,
there have been very few efforts to tackle such problems.
We present a novel end-to-end deep generative pipeline for improving classifier performance when dealing with such data problems. We call it
Uncertainty Mining Net (UMN).
During the training stage, we utilize all the available data (labeled and unlab eled) to train the classifier via a semi-supervised generative framework.
During training, UMN estimates the uncertainly of the labels' to focus on clean data for learning. More precisely, UMN applies the sample-wise label uncertain ty estimation scheme.
Extensive experiments and comparisons against state-of-the-art methods on sever al popular benchmark datasets demonstrate that UMN can reduce the effects of lab el noise and significantly improve classifier performance.
****
Information Geometry of Orthogonal Initializations and Training
Piotr Aleksander Sokó■,Il Memming Park
Recently mean field theory has been successfully used to analyze properties of wide, random neural networks. It gave rise to a prescriptive theory for initializing feed-forward neural networks with orthogonal weights, which ensures that both the forward propagated activations and the backpropagated gradients are near \(\ell_2\) isometries and as a consequence training is orders of magnitude faster. Despite strong empirical performance, the mechanisms by which critical initializations confer an advantage in the optimization of deep neural networks are poorly understood. Here we show a novel connection between the maximum curvature of the optimization landscape (gradient smoothness) as measured by the Fisher information matrix (FIM) and the spectral radius of the input-output Jacobian, which partially explains why more isometric networks can train much faster. Furthermore, given that orthogonal weights are necessary to ensure that gradient norms are approximately preserved at initialization, we experimentally investigate the benefits of maintaining orthogonality throughout training, and we conclude that manifold optimization of weights performs well regardless of the smoothness of the gradients. Moreover, we observe a surprising yet robust behavior of highly isometric initializations --- even though such networks have a lower FIM condition number \emph{at initialization}, and therefore by analogy to convex functions should be easier to optimize, experimentally they prove to be much harder to train with stochastic gradient descent. We conjecture the FIM condition number plays a non-trivial role in the optimiza tion.
****
Multi-Step Decentralized Domain Adaptation
Akhil Mathur,Shaoduo Gan,Anton Isopoussu,Fahim Kawsar,Nadia Berthouze,Nicholas D . Lane
Despite the recent breakthroughs in unsupervised domain adaptation (uDA), no pri or work has studied the challenges of applying these methods in practical machin e learning scenarios. In this paper, we highlight two significant bottlenecks fo r uDA, namely excessive centralization and poor support for distributed domain d atasets. Our proposed framework, MDDA, is powered by a novel collaborator select ion algorithm and an effective distributed adversarial training method, and allo ws for uDA methods to work in a decentralized and privacy-preserving way.

****
Mixed Precision DNNs: All you need is a good parametrization
Stefan Uhlich,Lukas Mauch,Fabien Cardinaux,Kazuki Yoshiyama,Javier Alonso Garcia ,Stephen Tiedemann,Thomas Kemp,Akira Nakamura
Efficient deep neural network (DNN) inference on mobile or embedded devices typi

cally involves quantization of the network parameters and activations. In partic
ular, mixed precision networks achieve better performance than networks with hom
ogeneous bitwidth for the same size constraint. Since choosing the optimal bitwi
dths is not straight forward, training methods, which can learn them, are desira
ble. Differentiable quantization with straight-through gradients allows to learn
 the quantizer's parameters using gradient methods. We show that a suited parame
trization of the quantizer is the key to achieve a stable training and a good fi
nal performance. Specifically, we propose to parametrize the quantizer with the
step size and dynamic range. The bitwidth can then be inferred from them. Other
parametrizations, which explicitly use the bitwidth, consistently perform worse.
 We confirm our findings with experiments on CIFAR-10 and ImageNet and we obtain
 mixed precision DNNs with learned quantization parameters, achieving state-of-t
he-art performance.
****

PROGRESSIVE LEARNING AND DISENTANGLEMENT OF HIERARCHICAL REPRESENTATIONS
Zhiyuan Li,Jaideep Vitthal Murkute,Prashnna Kumar Gyawali,Linwei Wang
Learning rich representation from data is an important task for deep generative
models such as variational auto-encoder (VAE). However, by extracting high-level
 abstractions in the bottom-up inference process, the goal of preserving all fac
tors of variations for top-down generation is compromised. Motivated by the conc
ept of “starting small”, we present a strategy to progressively learn independen
t hierarchical representations from high- to low-levels of abstractions. The mod
el starts with learning the most abstract representation, and then progressively
 grow the network architecture to introduce new representations at different le
vels of abstraction. We quantitatively demonstrate the ability of the presented
model to improve disentanglement in comparison to existing works on two benchmar
k datasets using three disentanglement metrics, including a new metric we propos
ed to complement the previously-presented metric of mutual information gap. We f
urther present both qualitative and quantitative evidence on how the progression
 of learning improves disentangling of hierarchical representations. By drawing
on the respective advantage of hierarchical representation learning and progress
ive learning, this is to our knowledge the first attempt to improve disentanglem
ent by progressively growing the capacity of VAE to learn hierarchical represent
ations.
****

Co-Attentive Equivariant Neural Networks: Focusing Equivariance On Transformatio
ns Co-Occurring in Data
David W. Romero,Mark Hoogendoorn
Equivariance is a nice property to have as it produces much more parameter effic
ient neural architectures and preserves the structure of the input through the f
eature mapping. Even though some combinations of transformations might never app
ear (e.g. an upright face with a horizontal nose), current equivariant architect
ures consider the set of all possible transformations in a transformation group
when learning feature representations. Contrarily, the human visual system is ab
le to attend to the set of relevant transformations occurring in the environment
 and utilizes this information to assist and improve object recognition. Based o
n this observation, we modify conventional equivariant feature mappings such tha
t they are able to attend to the set of co-occurring transformations in data and
 generalize this notion to act on groups consisting of multiple symmetries. We s
how that our proposed co-attentive equivariant neural networks consistently outp
erform conventional rotation equivariant and rotation & reflection equivariant n
eural networks on rotated MNIST and CIFAR-10.
****

Improving the Gating Mechanism of Recurrent Neural Networks
Albert Gua,Caglar Gulcehre,Tom le Paine,Razvan Pascanu,Matt Hoffman
In this work, we revisit the gating mechanisms widely used in various recurrent
and feedforward networks such as LSTMs, GRUs, or highway networks. These gates a
re meant to control information flow, allowing gradients to better propagate bac
k in time for recurrent models. However, to propagate gradients over very long t
emporal windows, they need to operate close to their saturation regime. We propo

se two independent and synergistic modifications to the standard gating mechanis m that are easy to implement, introduce no additional hyper-parameters, and are aimed at improving learnability of the gates when they are close to saturation. Our proposals are theoretically justified, and we show a generic framework that encompasses other recently proposed gating mechanisms such as chrono-initializat ion and master gates . We perform systematic analyses and ablation studies on th e proposed improvements and evaluate our method on a wide range of applications including synthetic memorization tasks, sequential image classification, languag e modeling, and reinforcement learning. Empirically, our proposed gating mechan isms robustly increase the performance of recurrent models such as LSTMs, especi ally on tasks requiring long temporal dependencies.
****

Learning to Transfer via Modelling Multi-level Task Dependency
Haonan Wang,Zhenbang Wu,Ziniu Hu,Yizhou Sun
Multi-task learning has been successful in modeling multiple related tasks with large, carefully curated labeled datasets. By leveraging the relationships among different tasks, multi-task learning framework can improve the performance sign ificantly. However, most of the existing works are under the assumption that the predefined tasks are related to each other. Thus, their applications on real-wo rld are limited, because rare real-world problems are closely related. Besides, the understanding of relationships among tasks has been ignored by most of the c urrent methods. Along this line, we propose a novel multi-task learning framewor k - Learning To Transfer Via Modelling Multi-level Task Dependency, which constr ucted attention based dependency relationships among different tasks. At the sam e time, the dependency relationship can be used to guide what knowledge should b e transferred, thus the performance of our model also be improved. To show the e ffectiveness of our model and the importance of considering multi-level dependen cy relationship, we conduct experiments on several public datasets, on which we obtain significant improvements over current methods.
****

Latent Variables on Spheres for Sampling and Inference
Deli Zhao,Jiapeng Zhu,Bo Zhang
Variational inference is a fundamental problem in Variational AutoEncoder (VAE). The optimization with lower bound of marginal log-likelihood results in the dis tribution of latent variables approximate to a given prior probability, which is the dilemma of employing VAE to solve real-world problems. By virtue of high-di mensional geometry, we propose a very simple algorithm completely different from existing ones to alleviate the variational inference in VAE. We analyze the uni que characteristics of random variables on spheres in high dimensions and prove that Wasserstein distance between two arbitrary data sets randomly drawn from a sphere are nearly identical when the dimension is sufficiently large. Based on o ur theory, a novel algorithm for distribution-robust sampling is devised. Moreov er, we reform the latent space of VAE by constraining latent variables on the sp here, thus freeing VAE from the approximate optimization of posterior probabilit y via variational inference. The new algorithm is named Spherical AutoEncoder (S AE). Extensive experiments by sampling and inference tasks validate our theoreti cal analysis and the superiority of SAE.
****

Deep Orientation Uncertainty Learning based on a Bingham Loss
Igor Gilitschenski,Roshni Sahoo,Wilko Schwarting,Alexander Amini,Sertac Karaman, Daniela Rus
Reasoning about uncertain orientations is one of the core problems in many perce ption tasks such as object pose estimation or motion estimation. In these scenar ios, poor illumination conditions, sensor limitations, or appearance invariance may result in highly uncertain estimates. In this work, we propose a novel learn ing-based representation for orientation uncertainty. By characterizing uncertai nty over unit quaternions with the Bingham distribution, we formulate a loss tha t naturally captures the antipodal symmetry of the representation. We discuss th e interpretability of the learned distribution parameters and demonstrate the fe asibility of our approach on several challenging real-world pose estimation task

s involving uncertain orientations.
****
Analyzing Privacy Loss in Updates of Natural Language Models
Shruti Tople,Marc Brockschmidt,Boris Köpf,Olga Ohrimenko,Santiago Zanella-Béguel in
To continuously improve quality and reflect changes in data, machine learning-ba sed services have to regularly re-train and update their core models. In the set ting of language models, we show that a comparative analysis of model snapshots before and after an update can reveal a surprising amount of detailed informatio n about the changes in the data used for training before and after the update. We discuss the privacy implications of our findings, propose mitigation strategi es and evaluate their effect.
****
Learning from Positive and Unlabeled Data with Adversarial Training
Wenpeng Hu,Ran Le,Bing Liu,Feng Ji,Haiqing Chen,Dongyan Zhao,Jinwen Ma,Rui Yan
Positive-unlabeled (PU) learning learns a binary classifier using only positive and unlabeled examples without labeled negative examples. This paper shows that the GAN (Generative Adversarial Networks) style of adversarial training is quite suitable for PU learning. GAN learns a generator to generate data (e.g., images) to fool a discriminator which tries to determine whether the generated data be long to a (positive) training class. PU learning is similar and can be naturally casted as trying to identify (not generate) likely positive data from the unlab eled set also to fool a discriminator that determines whether the identified lik ely positive data from the unlabeled set (U) are indeed positive (P). A direct a daptation of GAN for PU learning does not produce a strong classifier. This pape r proposes a more effective method called Predictive Adversarial Networks (PAN) using a new objective function based on KL-divergence, which performs much bette r.~Empirical evaluation using both image and text data shows the effectiveness o f PAN.
****
Deep exploration by novelty-pursuit with maximum state entropy
Zi-Niu Li,Xiong-Hui Chen,Yang Yu
Efficient exploration is essential to reinforcement learning in huge state space . Recent approaches to address this issue include the intrinsically motivated go al exploration process (IMGEP) and the maximum state entropy exploration (MSEE). In this paper, we disclose that goal-conditioned exploration behaviors in IMGEP can also maximize the state entropy, which bridges the IMGEP and the MSEE. From this connection, we propose a maximum entropy criterion for goal selection in g oal-conditioned exploration, which results in the new exploration method novelty -pursuit. Novelty-pursuit performs the exploration in two stages: first, it sele cts a goal for the goal-conditioned exploration policy to reach the boundary of the explored region; then, it takes random actions to explore the non-explored r egion. We demonstrate the effectiveness of the proposed method in environments f rom simple maze environments, Mujoco tasks, to the long-horizon video game of Su perMarioBros. Experiment results show that the proposed method outperforms the s tate-of-the-art approaches that use curiosity-driven exploration.
****
Reconstructing continuous distributions of 3D protein structure from cryo-EM ima ges
Ellen D. Zhong,Tristan Bepler,Joseph H. Davis,Bonnie Berger
Cryo-electron microscopy (cryo-EM) is a powerful technique for determining the s tructure of proteins and other macromolecular complexes at near-atomic resolutio n. In single particle cryo-EM, the central problem is to reconstruct the 3D stru cture of a macromolecule from 10^{4-7} noisy and randomly oriented 2D projecti on images. However, the imaged protein complexes may exhibit structural variabil ity, which complicates reconstruction and is typically addressed using discrete clustering approaches that fail to capture the full range of protein dynamics. H ere, we introduce a novel method for cryo-EM reconstruction that extends natural ly to modeling continuous generative factors of structural heterogeneity. This m ethod encodes structures in Fourier space using coordinate-based deep neural net

works, and trains these networks from unlabeled 2D cryo-EM images by combining e
xact inference over image orientation with variational inference for structural
heterogeneity. We demonstrate that the proposed method, termed cryoDRGN, can per
form ab-initio reconstruction of 3D protein complexes from simulated and real 2D
 cryo-EM image data. To our knowledge, cryoDRGN is the first neural network-base
d approach for cryo-EM reconstruction and the first end-to-end method for direct
ly reconstructing continuous ensembles of protein structures from cryo-EM images
.
****

Deep Evidential Uncertainty
Alexander Amini,Wilko Schwarting,Ava Soleimany,Daniela Rus
Deterministic neural networks (NNs) are increasingly being deployed in safety cr
itical domains, where calibrated, robust and efficient measures of uncertainty a
re crucial. While it is possible to train regression networks to output the para
meters of a probability distribution by maximizing a Gaussian likelihood functio
n, the resulting model remains oblivious to the underlying confidence of its pre
dictions. In this paper, we propose a novel method for training deterministic NN
s to not only estimate the desired target but also the associated evidence in su
pport of that target. We accomplish this by placing evidential priors over our
original Gaussian likelihood function and training our NN to infer the hyperpara
meters of our evidential distribution. We impose priors during training such tha
t the model is penalized when its predicted evidence is not aligned with the cor
rect output. Thus the model estimates not only the probabilistic mean and varian
ce of our target but also the underlying uncertainty associated with each of tho
se parameters. We observe that our evidential regression method learns well-cali
brated measures of uncertainty on various benchmarks, scales to complex computer
 vision tasks, and is robust to adversarial input perturbations.

****

Generalization of Two-layer Neural Networks: An Asymptotic Viewpoint
Jimmy Ba,Murat Erdogdu,Taiji Suzuki,Denny Wu,Tianzong Zhang
This paper investigates the generalization properties of two-layer neural networ
ks in high-dimensions, i.e. when the number of samples n, features d, and ne
urons h tend to infinity at the same rate. Specifically, we derive the exact p
opulation risk of the unregularized least squares regression problem with two-la
yer neural networks when either the first or the second layer is trained using a
 gradient flow under different initialization setups. When only the second laye
r coefficients are optimized, we recover the \textit{double descent} phenomenon:
 a cusp in the population risk appears at $h\approx n$ and further overparameter
ization decreases the risk. In contrast, when the first layer weights are optimi
zed, we highlight how different scales of initialization lead to different induc
tive bias, and show that the resulting risk is \textit{independent} of overparam
eterization. Our theoretical and experimental results suggest that previously st
udied model setups that provably give rise to \textit{double descent} might not
translate to optimizing two-layer neural networks.
****

Better Knowledge Retention through Metric Learning
Ke Li*,Shichong Peng*,Kailas Vodrahalli*,Jitendra Malik
In a continual learning setting, new categories may be introduced over time, and
 an ideal learning system should perform well on both the original categories an
d the new categories. While deep neural nets have achieved resounding success in
 the classical setting, they are known to forget about knowledge acquired in pri
or episodes of learning if the examples encountered in the current episode of le
arning are drastically different from those encountered in prior episodes. This
makes deep neural nets ill-suited to continual learning. In this paper, we propo
se a new model that can both leverage the expressive power of deep neural nets a
nd is resilient to forgetting when new categories are introduced. We demonstrate
 an improvement in terms of accuracy on original classes compared to a vanilla d
eep neural net.
****

Winning the Lottery with Continuous Sparsification
Pedro Savarese,Hugo Silva,Michael Maire
The Lottery Ticket Hypothesis from Frankle & Carbin (2019) conjectures that, for typically-sized neural networks, it is possible to find small sub-networks which train faster and yield superior performance than their original counterparts. The proposed algorithm to search for such sub-networks (winning tickets), Iterative Magnitude Pruning (IMP), consistently finds sub-networks with 90-95% less parameters which indeed train faster and better than the overparameterized models they were extracted from, creating potential applications to problems such as transfer learning.

In this paper, we propose a new algorithm to search for winning tickets, Continuous Sparsification, which continuously removes parameters from a network during training, and learns the sub-network's structure with gradient-based methods instead of relying on pruning strategies. We show empirically that our method is capable of finding tickets that outperforms the ones learned by Iterative Magnitude Pruning, and at the same time providing up to 5 times faster search, when measured in number of training epochs.
****
Critical initialisation in continuous approximations of binary neural networks
George Stamatescu,Federica Gerace,Carlo Lucibello,Ian Fuss,Langford White
The training of stochastic neural network models with binary (±1) weights and activations via continuous surrogate networks is investigated. We derive new surrogates using a novel derivation based on writing the stochastic neural network as a Markov chain. This derivation also encompasses existing variants of the surrogates presented in the literature. Following this, we theoretically study the surrogates at initialisation. We derive, using mean field theory, a set of scalar equations describing how input signals propagate through the randomly initialised networks. The equations reveal whether so-called critical initialisations exist for each surrogate network, where the network can be trained to arbitrary depth. Moreover, we predict theoretically and confirm numerically, that common weight initialisation schemes used in standard continuous networks, when applied to the mean values of the stochastic binary weights, yield poor training performance. This study shows that, contrary to common intuition, the means of the stochastic binary weights should be initialised close to ± 1, for deeper networks to be trainable.
****
LEARNING DIFFICULT PERCEPTUAL TASKS WITH HODGKIN-HUXLEY NETWORKS
Alan Lockett,Ankit Patel,Paul Pfaffinger
This paper demonstrates that a computational neural network model using ion channel-based conductances to transmit information can solve standard computer vision datasets at near state-of-the-art performance. Although not fully biologically accurate, this model incorporates fundamental biophysical principles underlying the control of membrane potential and the processing of information by Ohmic ion channels. The key computational step employs Conductance-Weighted Averaging (CWA) in place of the traditional affine transformation, representing a fundamentally different computational principle.
Importantly, CWA based networks are self-normalizing and range-limited. We also demonstrate for the first time that a network with excitatory and inhibitory neurons and nonnegative synapse strengths can successfully solve computer vision problems. Although CWA models do not yet surpass the current state-of-the-art in deep learning, the results are competitive on CIFAR-10. There remain avenues for improving these networks, e.g. by more closely modeling ion channel function and connectivity patterns of excitatory and inhibitory neurons found in the brain.

****
Filter redistribution templates for iteration-lessconvolutional model reduction
Ramon Izquierdo Cordova,Walterio Mayol Cuevas
Automatic neural network discovery methods face an enormous challenge caused for the size of the search space. A common practice is to split this space at diffe

rent levels and to explore only a part of it. Neural architecture search methods look for how to combine a subset of layers, which are the most promising, to create an architecture while keeping a predefined number of filters in each layer. On the other hand, pruning techniques take a well known architecture and look for the appropriate number of filters per layer. In both cases the exploration is made iteratively, training models several times during the search. Inspired by the advantages of the two previous approaches, we proposed a fast option to find models with improved characteristics. We apply a small set of templates, which are considered promising, for make a redistribution of the number of filters in an already existing neural network. When compared to the initial base models, we found that the resulting architectures, trained from scratch, surpass the original accuracy even after been reduced to fit the same amount of resources.
****

Universal Safeguarded Learned Convex Optimization with Guaranteed Convergence
Howard Heaton,Xiaohan Chen,Zhangyang Wang,Wotao Yin
Many applications require quickly and repeatedly solving a certain type of optimization problem, each time with new (but similar) data. However, state of the art general-purpose optimization methods may converge too slowly for real-time use. This shortcoming is addressed by "learning to optimize" (L2O) schemes, which construct neural networks from parameterized forms of the update operations of general-purpose methods. Inferences by each network form solution estimates, and networks are trained to optimize these estimates for a particular distribution of data. This results in task-specific algorithms (e.g., LISTA, ALISTA, and D-LADMM) that can converge order(s) of magnitude faster than general-purpose counterparts. We provide the first general L2O convergence theory by wrapping all L2O schemes for convex optimization within a single framework. Existing L2O schemes form special cases, and we give a practical guide for applying our L2O framework to other problems. Using safeguarding, our theory proves, as the number of network layers increases, the distance between inferences and the solution set goes to zero, i.e., each cluster point is a solution. Our numerical examples demonstrate the efficacy of our approach for both existing and new L2O methods.
****

A Gradient-Based Approach to Neural Networks Structure Learning
Amir Ali Moinfar,Amirkeivan Mohtashami,Mahdieh Soleymani,Ali Sharifi-Zarchi
Designing the architecture of deep neural networks (DNNs) requires human expertise and is a cumbersome task. One approach to automatize this task has been considering DNN architecture parameters such as the number of layers, the number of neurons per layer, or the activation function of each layer as hyper-parameters, and using an external method for optimizing it. Here we propose a novel neural network model, called Farfalle Neural Network, in which important architecture features such as the number of neurons in each layer and the wiring among the neurons are automatically learned during the training process. We show that the proposed model can replace a stack of dense layers, which is used as a part of many DNN architectures. It can achieve higher accuracy using significantly fewer parameters.
****

Sub-policy Adaptation for Hierarchical Reinforcement Learning
Alexander Li,Carlos Florensa,Ignasi Clavera,Pieter Abbeel
Hierarchical reinforcement learning is a promising approach to tackle long-horizon decision-making problems with sparse rewards. Unfortunately, most methods still decouple the lower-level skill acquisition process and the training of a higher level that controls the skills in a new task. Leaving the skills fixed can lead to significant sub-optimality in the transfer setting. In this work, we propose a novel algorithm to discover a set of skills, and continuously adapt them along with the higher level even when training on a new task. Our main contributions are two-fold. First, we derive a new hierarchical policy gradient with an unbiased latent-dependent baseline, and we introduce Hierarchical Proximal Policy Optimization (HiPPO), an on-policy method to efficiently train all levels of the hierarchy jointly. Second, we propose a method of training time-abstractions that improves the robustness of the obtained skills to environment changes. Code a

nd videos are available at sites.google.com/view/hippo-rl.
****
AdvCodec: Towards A Unified Framework for Adversarial Text Generation
Boxin Wang,Hengzhi Pei,Han Liu,Bo Li
Machine learning (ML) especially deep neural networks (DNNs) have been widely ap
plied to real-world applications. However, recent studies show that DNNs are vul
nerable to carefully crafted \emph{adversarial examples} which only deviate fro
m the original data by a small magnitude of perturbation.
While there has been great interest on generating imperceptible adversarial exam
ples in continuous data domain (e.g. image and audio) to explore the model vulne
rabilities, generating \emph{adversarial text} in the discrete domain is still c
hallenging.
The main contribution of this paper is to propose a general targeted attack fram
ework \advcodec for adversarial text generation which addresses the challenge of
 discrete input space and be easily adapted to general natural language processi
ng (NLP) tasks.
In particular, we propose a tree based autoencoder to encode discrete text data
into continuous vector space, upon which we optimize the adversarial perturbatio
n. With the tree based decoder, it is possible to ensure the grammar correctness
 of the generated text; and the tree based encoder enables flexibility of making
 manipulations on different levels of text, such as sentence (\advcodecsent) and
 word (\advcodecword) levels. We consider multiple attacking scenarios, includin
g appending an adversarial sentence or adding unnoticeable words to a given para
graph, to achieve arbitrary \emph{targeted attack}. To demonstrate the effective
ness of the proposed method, we consider two most representative NLP tasks: sent
iment analysis and question answering (QA). Extensive experimental results show
that \advcodec has successfully attacked both tasks. In particular, our attack c
auses a BERT-based sentiment classifier accuracy to drop from 0.703 to 0.006
, and a BERT-based QA model's F1 score to drop from 88.62 to 33.21 (with bes
t targeted attack F1 score as 46.54). Furthermore, we show that the white-box
generated adversarial texts can transfer across other black-box models, shedding
 light on an effective way to examine the robustness of existing NLP models.
****
PROVABLY BENEFITS OF DEEP HIERARCHICAL RL
Zeyu Jia,Simon S. Du,Ruosong Wang,Mengdi Wang,Lin F. Yang
Modern complex sequential decision-making problem often both low-level policy an
d high-level planning. Deep hierarchical reinforcement learning (Deep HRL) admit
s multi-layer abstractions which naturally model the policy in a hierarchical ma
nner, and it is believed that deep HRL can reduce the sample complexity compared
 to the standard RL frameworks. We initiate the study of rigorously characterizi
ng the complexity of Deep HRL. We present a model-based optimistic algorithm whi
ch demonstrates that the complexity of learning a near-optimal policy for deep H
RL scales with the sum of number of states at each abstraction layer whereas sta
ndard RL scales with the product of number of states at each abstraction layer.
Our algorithm achieves this goal by using the fact that distinct high-level stat
es have similar low-level structures, which allows an efficient information expl
oitation and thus experiences from different high-level state-action pairs can b
e generalized to unseen state-actions. Overall, our result shows an exponential
improvement using Deep HRL comparing to standard RL framework.
****
Learning Latent State Spaces for Planning through Reward Prediction
Aaron Havens,Yi Ouyang,Prabhat Nagarajan,Yasuhiro Fujita
Model-based reinforcement learning methods typically learn models for high-dimen
sional state spaces by aiming to reconstruct and predict the original observatio
ns. However, drawing inspiration from model-free reinforcement learning, we prop
ose learning a latent dynamics model directly from rewards. In this work, we int
roduce a model-based planning framework which learns a latent reward prediction
model and then plan in the latent state-space. The latent representation is lear
ned exclusively from multi-step reward prediction which we show to be the only n
ecessary information for successful planning. With this framework, we are able

to benefit from the concise model-free representation, while still enjoying the data-efficiency of model-based algorithms. We demonstrate our framework in multi-pendulum and multi-cheetah environments where several pendulums or cheetahs are shown to the agent but only one of them produces rewards. In these environments, it is important for the agent to construct a concise latent representation to filter out irrelevant observations. We find that our method can successfully learn an accurate latent reward prediction model in the presence of the irrelevant information while existing model-based methods fail. Planning in the learned latent state-space shows strong performance and high sample efficiency over model-free and model-based baselines.
****

Hope For The Best But Prepare For The Worst: Cautious Adaptation In RL Agents
Jesse Zhang,Brian Cheung,Chelsea Finn,Dinesh Jayaraman,Sergey Levine
We study the problem of safe adaptation: given a model trained on a variety of past experiences for some task, can this model learn to perform that task in a new situation while avoiding catastrophic failure? This problem setting occurs frequently in real-world reinforcement learning scenarios such as a vehicle adapting to drive in a new city, or a robotic drone adapting a policy trained only in simulation. While learning without catastrophic failures is exceptionally difficult, prior experience can allow us to learn models that make this much easier. These models might not directly transfer to new settings, but can enable cautious adaptation that is substantially safer than na\"{i}ve adaptation as well as learning from scratch. Building on this intuition, we propose risk-averse domain adaptation (RADA). RADA works in two steps: it first trains probabilistic model-based RL agents in a population of source domains to gain experience and capture epistemic uncertainty about the environment dynamics. Then, when dropped into a new environment, it employs a pessimistic exploration policy, selecting actions that have the best worst-case performance as forecasted by the probabilistic model. We show that this simple maximin policy accelerates domain adaptation in a safety-critical driving environment with varying vehicle sizes. We compare our approach against other approaches for adapting to new environments, including meta-reinforcement learning.
****

Semi-Supervised Boosting via Self Labelling
Akul Goyal,Yang Liu
Attention to semi-supervised learning grows in machine learning as the price to expertly label data increases. Like most previous works in the area, we focus on improving an algorithm's ability to discover the inherent property of the entire dataset from a few expertly labelled samples. In this paper we introduce Boosting via Self Labelling (BSL), a solution to semi-supervised boosting when there is only limited access to labelled instances. Our goal is to learn a classifier that is trained on a data set that is generated by combining the generalization of different algorithms which have been trained with a limited amount of supervised training samples. Our method builds upon a combination of several different components. First, an inference aided ensemble algorithm developed on a set of weak classifiers will offer the initial noisy labels. Second, an agreement based estimation approach will return the average error rates of the noisy labels. Third and finally, a noise-resistant boosting algorithm will train over the noisy labels and their error rates to describe the underlying structure as closely as possible. We provide both analytical justifications and experimental results to back the performance of our model. Based on several benchmark datasets, our results demonstrate that BSL is able to outperform state-of-the-art semi-supervised methods consistently, achieving over 90% test accuracy with only 10% of the data being labelled.
****

Fractional Graph Convolutional Networks (FGCN) for Semi-Supervised Learning
Yuzhou Chen,Yulia R. Gel,Konstantin Avrachenkov
Due to high utility in many applications, from social networks to blockchain to power grids, deep learning on non-Euclidean objects such as graphs and manifolds continues to gain an ever increasing interest. Most currently available techniq

ues are based on the idea of performing a convolution operation in the spectral domain with a suitably chosen nonlinear trainable filter and then approximating the filter with finite order polynomials. However, such polynomial approximation approaches tend to be both non-robust to changes in the graph structure and to capture primarily the global graph topology. In this paper we propose a new Fractional Generalized Graph Convolutional Networks (FGCN) method for semi-supervised learning, which casts the L\'evy Fights into random walks on graphs and, as a result, allows to more accurately account for the intrinsic graph topology and to substantially improve classification performance, especially for heterogeneous graphs.
****

Antifragile and Robust Heteroscedastic Bayesian Optimisation
Ryan Rhys-Griffiths,Miguel Garcia-Ortegon,Alexander A. Aldrick,Alpha A. Lee
    Bayesian Optimisation is an important decision-making tool for high-stakes applications in drug discovery and materials design. An oft-overlooked modelling consideration however is the representation of input-dependent or heteroscedastic aleatoric uncertainty. The cost of misrepresenting this uncertainty as being homoscedastic could be high in drug discovery applications where neglecting heteroscedasticity in high throughput virtual screening could lead to a failed drug discovery program. In this paper, we propose a heteroscedastic Bayesian Optimisation scheme which both represents and optimises aleatoric noise in the suggestions. We consider cases such as drug discovery where we would like to minimise or be robust to aleatoric uncertainty but also applications such as materials discovery where it may be beneficial to maximise or be antifragile to aleatoric uncertainty. Our scheme features a heteroscedastic Gaussian Process (GP) as the surrogate model in conjunction with two acquisition heuristics. First, we extend the augmented expected improvement (AEI) heuristic to the heteroscedastic setting and second, we introduce a new acquisition function, aleatoric-penalised expected improvement (ANPEI) based on a simple scalarisation of the performance and noise objective. Both methods are capable of penalising or promoting aleatoric noise in the suggestions and yield improved performance relative to a naive implementation of homoscedastic Bayesian Optimisation on toy problems as well as a real-world optimisation problem.
****

Generalizing Reinforcement Learning to Unseen Actions
Ayush Jain*,Andrew Szot*,Jincheng Zhou,Joseph J. Lim
A fundamental trait of intelligence is the ability to achieve goals in the face of novel circumstances. In this work, we address one such setting which requires solving a task with a novel set of actions. Empowering machines with this ability requires generalization in the way an agent perceives its available actions along with the way it uses these actions to solve tasks. Hence, we propose a framework to enable generalization over both these aspects: understanding an action’s functionality, and using actions to solve tasks through reinforcement learning. Specifically, an agent interprets an action’s behavior using unsupervised representation learning over a collection of data samples reflecting the diverse properties of that action. We employ a reinforcement learning architecture which works over these action representations, and propose regularization metrics essential for enabling generalization in a policy. We illustrate the generalizability of the representation learning method and policy, to enable zero-shot generalization to previously unseen actions on challenging sequential decision-making environments. Our results and videos can be found at sites.google.com/view/action-generalization/
****

Provable Representation Learning for Imitation Learning via Bi-level Optimization
Sanjeev Arora,Simon S. Du,Sham Kakade,Yuping Luo,Nikunj Saunshi
A common strategy in modern learning systems is to learn a representation which is useful for many tasks, a.k.a, representation learning. We study this strategy in the imitation learning setting where multiple experts trajectories are available. We formulate representation learning as a bi-level optimization problem wh

ere the "outer" optimization tries to learn the joint representation and the "in
ner" optimization encodes the imitation learning setup and tries to learn task-s
pecific parameters. We instantiate this framework for the cases where the imitat
ion setting being behavior cloning and observation alone. Theoretically, we prov
ably show using our framework that representation learning can reduce the sample
 complexity of imitation learning in both settings. We also provide proof-of-con
cept experiments to verify our theoretical findings.
****

Episodic Reinforcement Learning with Associative Memory
Guangxiang Zhu*,Zichuan Lin*,Guangwen Yang,Chongjie Zhang
Sample efficiency has been one of the major challenges for deep reinforcement le
arning. Non-parametric episodic control has been proposed to speed up parametric
 reinforcement learning by rapidly latching on previously successful policies. H
owever, previous work on episodic reinforcement learning neglects the relationsh
ip between states and only stored the experiences as unrelated items. To improve
 sample efficiency of reinforcement learning, we propose a novel framework, call
ed Episodic Reinforcement Learning with Associative Memory (ERLAM), which associ
ates related experience trajectories to enable reasoning effective strategies. W
e build a graph on top of states in memory based on state transitions and develo
p a reverse-trajectory propagation strategy to allow rapid value propagation thr
ough the graph. We use the non-parametric associative memory as early guidance f
or a parametric reinforcement learning model. Results on navigation domain and A
tari games show our framework achieves significantly higher sample efficiency th
an state-of-the-art episodic reinforcement learning models.
****

Flexible and Efficient Long-Range Planning Through Curious Exploration
Aidan Curtis,Minjian Xin,Kevin Feigelis,Dan Yamins
Identifying algorithms that flexibly and efficiently discover temporally-extende
d multi-phase plans is an essential next step for the advancement of robotics an
d model-based reinforcement learning. The core problem of long-range planning is
 finding an efficient way to search through the tree of possible action sequence
s — which, if left unchecked, grows exponentially with the length of the plan. E
xisting non-learned planning solutions from the Task and Motion Planning (TAMP)
literature rely on the existence of logical descriptions for the effects and pre
conditions for actions. This constraint allows TAMP methods to efficiently reduc
e the tree search problem but limits their ability to generalize to unseen and c
omplex physical environments. In contrast, deep reinforcement learning (DRL) met
hods use flexible neural-network-based function approximators to discover polici
es that generalize naturally to unseen circumstances. However, DRL methods have
had trouble dealing with the very sparse reward landscapes inherent to long-rang
e multi-step planning situations. Here, we propose the Curious Sample Planner (C
SP), which fuses elements of TAMP and DRL by using a curiosity-guided sampling s
trategy to learn to efficiently explore the tree of action effects. We show that
 CSP can efficiently discover interesting and complex temporally-extended plans
for solving a wide range of physically realistic 3D tasks. In contrast, standard
 DRL and random sampling methods often fail to solve these tasks at all or do so
 only with a huge and highly variable number of training samples. We explore the
 use of a variety of curiosity metrics with CSP and analyze the types of solutio
ns that CSP discovers. Finally, we show that CSP supports task transfer so that
the exploration policies learned during experience with one task can help improv
e efficiency on related tasks.
****

Learning to Prove Theorems by Learning to Generate Theorems
Mingzhe Wang,Jia Deng
We consider the task of automated theorem proving, a key AI task. Deep learning
has shown promise for training theorem provers, but there are limited human-writ
ten theorems and proofs available for supervised learning. To address this limit
ation, we propose to learn a neural generator that automatically synthesizes the
orems and proofs for the purpose of training a theorem prover. Experiments on re
al-world tasks demonstrate that synthetic data from our approach significantly

improves the theorem prover and advances the state of the art of automated theor
em proving in Metamath.
****
Depth-Width Trade-offs for ReLU Networks via Sharkovsky's Theorem
Vaggos Chatziafratis,Sai Ganesh Nagarajan,Ioannis Panageas,Xiao Wang
Understanding the representational power of Deep Neural Networks (DNNs) and how
their structural properties (e.g., depth, width, type of activation unit) affect
 the functions they can compute, has been an important yet challenging question
in deep learning and approximation theory. In a seminal paper, Telgarsky high- l
ighted the benefits of depth by presenting a family of functions (based on sim-
ple triangular waves) for which DNNs achieve zero classification error, whereas
shallow networks with fewer than exponentially many nodes incur constant error.
Even though Telgarsky's work reveals the limitations of shallow neural networks,
 it doesn't inform us on why these functions are difficult to represent and in f
act he states it as a tantalizing open question to characterize those functions
that cannot be well-approximated by smaller depths.
In this work, we point to a new connection between DNNs expressivity and Sharkov
sky's Theorem from dynamical systems, that enables us to characterize the depth-
width trade-offs of ReLU networks for representing functions based on the presen
ce of a generalized notion of fixed points, called periodic points (a fixed poin
t is a point of period 1). Motivated by our observation that the triangle waves
used in Telgarsky's work contain points of period 3 – a period that is special i
n that it implies chaotic behaviour based on the celebrated result by Li-Yorke –
 we proceed to give general lower bounds for the width needed to represent perio
dic functions as a function of the depth. Technically, the crux of our approach
is based on an eigenvalue analysis of the dynamical systems associated with such
 functions.
****
Gradient-based training of Gaussian Mixture Models in High-Dimensional Spaces
Alexander Gepperth,Benedikt Pfülb
We present an approach for efficiently training Gaussian Mixture Models (GMMs) w
ith Stochastic Gradient Descent (SGD) on large amounts of high-dimensional data
(e.g., images). In such a scenario, SGD is strongly superior in terms of executi
on time and memory usage, although it is conceptually more complex than the trad
itional Expectation-Maximization (EM) algorithm.
For enabling SGD training, we propose three novel ideas:
First, we show that minimizing an upper bound to the GMM log likelihood instead
of the full one is feasible and numerically much more stable way in high-dimensi
onal spaces.
Secondly, we propose a new regularizer that prevents SGD from converging to path
ological local minima.
And lastly, we present a simple method for enforcing the constraints inherent to
 GMM training when using SGD.
We also propose an SGD-compatible simplification to the full GMM model based on
local principal directions, which avoids excessive memory use in high-dimensiona
l spaces due to quadratic growth of covariance matrices.
Experiments on several standard image datasets show the validity of our approach
, and we provide a publicly available TensorFlow implementation.
****
Neural Phrase-to-Phrase Machine Translation
Jiangtao,Feng,Lingpeng Kong,Po-sen Huang,Chong,Wang,Da,Huang Jiayuan,Mao,Kan,Qia
o,Dengyong,Zhou
We present Neural Phrase-to-Phrase Machine Translation (\nppmt), a phrase-based
translation model that uses a novel phrase-attention mechanism to discover relev
ant input (source) segments to generate output (target) phrases. We propose an e
fficient dynamic programming algorithm to marginalize over all possible segments
 at training time and use a greedy algorithm or beam search for decoding. We als
o show how to incorporate a memory module derived from an external phrase dictio
nary to \nppmt{} to improve decoding. %that allows %the model to be trained fast
er %\nppmt is significantly faster %than existing neural phrase-based %machine t

ranslation method by \cite{huang2018towards}. Experiment results demonstrate tha
t \nppmt{} outperforms the best neural phrase-based translation model \citep{hua
ng2018towards} both in terms of model performance and speed, and is comparable t
o a state-of-the-art Transformer-based machine translation system \citep{vaswani
2017attention}.
****
At Your Fingertips: Automatic Piano Fingering Detection
Amit Moryossef,Yanai Elazar,Yoav Goldberg
Automatic Piano Fingering is a hard task which computers can learn using data. A
s data collection is hard and expensive, we propose to automate this process by
automatically extracting fingerings from public videos and MIDI files, using com
puter-vision techniques. Running this process on 90 videos results in the larges
t dataset for piano fingering with more than 150K notes. We show that when runni
ng a previously proposed model for automatic piano fingering on our dataset and
then fine-tuning it on manually labeled piano fingering data, we achieve state-o
f-the-art results.
In addition to the fingering extraction method, we also introduce a novel method
 for transferring deep-learning computer-vision models to work on out-of-domain
data, by fine-tuning it on out-of-domain augmentation proposed by a Generative A
dversarial Network (GAN).

For demonstration, we anonymously release a visualization of the output of our p
rocess for a single video on https://youtu.be/Gfs1UWQhr5Q
****
Energy-based models for atomic-resolution protein conformations
Yilun Du,Joshua Meier,Jerry Ma,Rob Fergus,Alexander Rives
We propose an energy-based model (EBM) of protein conformations that operates at
 atomic scale. The model is trained solely on crystallized protein data. By cont
rast, existing approaches for scoring conformations use energy functions that in
corporate knowledge of physical principles and features that are the complex pro
duct of several decades of research and tuning. To evaluate the model, we benchm
ark on the rotamer recovery task, the problem of predicting the conformation of
a side chain from its context within a protein structure, which has been used to
 evaluate energy functions for protein design. The model achieves performance cl
ose to that of the Rosetta energy function, a state-of-the-art method widely use
d in protein structure prediction and design. An investigation of the model’s ou
tputs and hidden representations finds that it captures physicochemical properti
es relevant to protein energy.
****
Federated Learning with Matched Averaging
Hongyi Wang,Mikhail Yurochkin,Yuekai Sun,Dimitris Papailiopoulos,Yasaman Khazaen
i
Federated learning allows edge devices to collaboratively learn a shared model w
hile keeping the training data on device, decoupling the ability to do model tra
ining from the need to store the data in the cloud. We propose Federated matched
 averaging (FedMA) algorithm designed for federated learning of modern neural ne
twork architectures e.g. convolutional neural networks (CNNs) and LSTMs. FedMA c
onstructs the shared global model in a layer-wise manner by matching and averagi
ng hidden elements (i.e. channels for convolution layers; hidden states for LSTM
; neurons for fully connected layers) with similar feature extraction signatures
. Our experiments indicate that FedMA not only outperforms popular state-of-the-
art federated learning algorithms on deep CNN and LSTM architectures trained on
real world datasets, but also reduces the overall communication burden.
****
Clustered Reinforcement Learning
Xiao Ma,Shen-Yi Zhao,Zhao-Heng Yin,Wu-Jun Li
Exploration strategy design is one of the challenging problems in reinforcement
learning~(RL), especially when the environment contains a large state space or s
parse rewards. During exploration, the agent tries to discover novel areas or hi
gh reward~(quality) areas. In most existing methods, the novelty and quality in

the neighboring area of the current state are not well utilized to guide the exp
loration of the agent. To tackle this problem, we propose a novel RL framework,
called \underline{c}lustered \underline{r}einforcement \underline{l}earning~(CRL
), for efficient exploration in RL. CRL adopts clustering to divide the collecte
d states into several clusters, based on which a bonus reward reflecting both no
velty and quality in the neighboring area~(cluster) of the current state is give
n to the agent. Experiments on several continuous control tasks and several Atar
i-2600 games show that CRL can outperform other state-of-the-art methods to achi
eve the best performance in most cases.
****

Pitfalls of In-Domain Uncertainty Estimation and Ensembling in Deep Learning
Arsenii Ashukha,Alexander Lyzhov,Dmitry Molchanov,Dmitry Vetrov
Uncertainty estimation and ensembling methods go hand-in-hand. Uncertainty estim
ation is one of the main benchmarks for assessment of ensembling performance. At
 the same time, deep learning ensembles have provided state-of-the-art results i
n uncertainty estimation. In this work, we focus on in-domain uncertainty for im
age classification. We explore the standards for its quantification and point ou
t pitfalls of existing metrics. Avoiding these pitfalls, we perform a broad stud
y of different ensembling techniques. To provide more insight in this study, we
introduce the deep ensemble equivalent score (DEE) and show that many sophistica
ted ensembling techniques are equivalent to an ensemble of only few independentl
y trained networks in terms of test performance.
****

Handwritten Amharic Character Recognition System Using Convolutional Neural Netw
orks
Fetulhak Abdurahman
Amharic language is an official language of the federal government of the Federa
l Democratic Republic of Ethiopia. Accordingly, there is a bulk of handwritten A
mharic documents available in libraries, information centres, museums, and offic
es. Digitization of these documents enables to harness already available languag
e technologies to local information needs and developments. Converting these doc
uments will have a lot of advantages including (i) to preserve and transfer hist
ory of the country (ii) to save storage space (ii) proper handling of documents
(iv) enhance retrieval of information through internet and other applications. H
andwritten Amharic character recognition system becomes a challenging task due t
o inconsistency of a writer, variability in writing styles of different writers,
 relatively large number of characters of the script, high interclass similarity
, structural complexity and degradation of documents due to different reasons. I
n order to recognize handwritten Amharic character a novel method based on deep
neural networks is used which has recently shown exceptional performance in vari
ous pattern recognition and machine learning applications, but has not been ende
avoured for Ethiopic script. The CNN model is trained and tested our database th
at contains 132,500 datasets of handwritten Amharic characters. Common machine l
earning methods usually apply a combination of feature extractor and trainable c
lassifier. The use of CNN leads to significant improvements across different mac
hine-learning classification algorithms. Our proposed CNN model is giving an acc
uracy of 91.83% on training data and 90.47% on validation data.
****

Effects of Linguistic Labels on Learned Visual Representations in Convolutional
Neural Networks: Labels matter!
Seoyoung Ahn,Gregory Zelinsky,Gary Lupyan
We investigated the changes in visual representations learnt by CNNs when using
different linguistic labels (e.g., trained with basic-level labels only, superor
dinate-level only, or both at the same time) and how they compare to human behav
ior when asked to select which of three images is most different. We compared CN
Ns with identical architecture and input, differing only in what labels were use
d to supervise the training. The results showed that in the absence of labels, t
he models learn very little categorical structure that is often assumed to be in
 the input. Models trained with superordinate labels (vehicle, tool, etc.) are m
ost helpful in allowing the models to match human categorization, implying that

human representations used in odd-one-out tasks are highly modulated by semantic information not obviously present in the visual input.
****
DiffTaichi: Differentiable Programming for Physical Simulation
Yuanming Hu,Luke Anderson,Tzu-Mao Li,Qi Sun,Nathan Carr,Jonathan Ragan-Kelley,Fredo Durand
We present DiffTaichi, a new differentiable programming language tailored for building high-performance differentiable physical simulators. Based on an imperative programming language, DiffTaichi generates gradients of simulation steps using source code transformations that preserve arithmetic intensity and parallelism. A light-weight tape is used to record the whole simulation program structure and replay the gradient kernels in a reversed order, for end-to-end backpropagation.
We demonstrate the performance and productivity of our language in gradient-based learning and optimization tasks on 10 different physical simulators. For example, a differentiable elastic object simulator written in our language is 4.2x shorter than the hand-engineered CUDA version yet runs as fast, and is 188x faster than the TensorFlow implementation.
Using our differentiable programs, neural network controllers are typically optimized within only tens of iterations.
****
Implicit Rugosity Regularization via Data Augmentation
Daniel LeJeune,Randall Balestriero,Hamid Javadi,Richard G. Baraniuk
Deep (neural) networks have been applied productively in a wide range of supervised and unsupervised learning tasks. Unlike classical machine learning algorithms, deep networks typically operate in the overparameterized regime, where the number of parameters is larger than the number of training data points. Consequently, understanding the generalization properties and the role of (explicit or implicit) regularization in these networks is of great importance. In this work, we explore how the oft-used heuristic of data augmentation imposes an implicit regularization penalty of a novel measure of the rugosity or “roughness” based on the tangent Hessian of the function fit to the training data.
****
A Mutual Information Maximization Perspective of Language Representation Learning
Lingpeng Kong,Cyprien de Masson d'Autume,Lei Yu,Wang Ling,Zihang Dai,Dani Yogatama
We show state-of-the-art word representation learning methods maximize an objective function that is a lower bound on the mutual information between different parts of a word sequence (i.e., a sentence). Our formulation provides an alternative perspective that unifies classical word embedding models (e.g., Skip-gram) and modern contextual embeddings (e.g., BERT, XLNet). In addition to enhancing our theoretical understanding of these methods, our derivation leads to a principled framework that can be used to construct new self-supervised tasks. We provide an example by drawing inspirations from related methods based on mutual information maximization that have been successful in computer vision, and introduce a simple self-supervised objective that maximizes the mutual information between a global sentence representation and n-grams in the sentence. Our analysis offers a holistic view of representation learning methods to transfer knowledge and translate progress across multiple domains (e.g., natural language processing, computer vision, audio processing).
****
Goal-Conditioned Video Prediction
Oleh Rybkin,Karl Pertsch,Frederik Ebert,Dinesh Jayaraman,Chelsea Finn,Sergey Levine
Many processes can be concisely represented as a sequence of events leading from a starting state to an end state. Given raw ingredients, and a finished cake, an experienced chef can surmise the recipe. Building upon this intuition, we propose a new class of visual generative models: goal-conditioned predictors (GCP). Prior work on video generation largely focuses on prediction models that only o

bserve frames from the beginning of the video. GCP instead treats videos as star t-goal transformations, making video generation easier by conditioning on the mo re informative context provided by the first and final frames. Not only do exis ting forward prediction approaches synthesize better and longer videos when modi fied to become goal-conditioned, but GCP models can also utilize structures tha t are not linear in time, to accomplish hierarchical prediction. To this end, we study both auto-regressive GCP models and novel tree-structured GCP models that generate frames recursively, splitting the video iteratively into finer and fin er segments delineated by subgoals. In experiments across simulated and real dat asets, our GCP methods generate high-quality sequences over long horizons. Tree -structured GCPs are also substantially easier to parallelize than auto-regressi ve GCPs, making training and inference very efficient, and allowing the mode l to train on sequences that are thousands of frames in length.Finally, we demon strate the utility of GCP approaches for imitation learning in the setting witho ut access to expert actions. Videos are on the supplementary website: https://si tes.google.com/view/video-gcp

****

Compression without Quantization

Gergely Flamich,Marton Havasi,José Miguel Hernández-Lobato

 Standard compression algorithms work by mapping an image to discrete code using an encoder from which the original image can be reconstructed through a decoder . This process, due to the quantization step, is inherently non-differentiable s o these algorithms must rely on approximate methods to train the encoder and dec oder end-to-end. In this paper, we present an innovative framework for lossy ima ge compression which is able to circumvent the quantization step by relying on a non-deterministic compression codec. The decoder maps the input image to a dist ribution in continuous space from which a sample can be encoded with expected co de length being the relative entropy to the encoding distribution, i.e. it is bi ts-back efficient. The result is a principled, end-to-end differentiable compres sion framework that can be straight-forwardly trained using standard gradient-ba sed optimizers. To showcase the efficiency of our method, we apply it to lossy i mage compression by training Probabilistic Ladder Networks (PLNs) on the CLIC 20 18 dataset and show that their rate-distortion curves on the Kodak dataset are c ompetitive with the state-of-the-art on low bitrates.

****

A Non-asymptotic comparison of SVRG and SGD: tradeoffs between compute and speed

Qingru Zhang,Yuhuai Wu,Fartash Faghri,Tianzong Zhang,Jimmy Ba

Stochastic gradient descent (SGD), which trades off noisy gradient updates for computational efficiency, is the de-facto optimization algorithm to solve large- scale machine learning problems. SGD can make rapid learning progress by perform ing updates using subsampled training data, but the noisy updates also lead to s low asymptotic convergence. Several variance reduction algorithms, such as SVR G, introduce control variates to obtain a lower variance gradient estimate and f aster convergence. Despite their appealing asymptotic guarantees, SVRG-like alg orithms have not been widely adopted in deep learning. The traditional asymptoti c analysis in stochastic optimization provides limited insight into training dee p learning models under a fixed number of epochs. In this paper, we present a no n-asymptotic analysis of SVRG under a noisy least squares regression problem. Ou r primary focus is to compare the exact loss of SVRG to that of SGD at each iter ation t. We show that the learning dynamics of our regression model closely matc hes with that of neural networks on MNIST and CIFAR-10 for both the underparamet erized and the overparameterized models. Our analysis and experimental results s uggest there is a trade-off between the computational cost and the convergence s peed in underparametrized neural networks. SVRG outperforms SGD after a few epoc hs in this regime. However, SGD is shown to always outperform SVRG in the overpa rameterized regime.

****

Towards Understanding the Spectral Bias of Deep Learning

Yuan Cao,Zhiying Fang,Yue Wu,Ding-Xuan Zhou,Quanquan Gu

An intriguing phenomenon observed during training neural networks is the spectra

l bias, where neural networks are biased towards learning less complex functions. The priority of learning functions with low complexity might be at the core of explaining generalization ability of neural network, and certain efforts have been made to provide theoretical explanation for spectral bias. However, there is still no satisfying theoretical results justifying the existence of spectral bias. In this work, we give a comprehensive and rigorous explanation for spectral bias and relate it with the neural tangent kernel function proposed in recent work. We prove that the training process of neural networks can be decomposed along different directions defined by the eigenfunctions of the neural tangent kernel, where each direction has its own convergence rate and the rate is determined by the corresponding eigenvalue. We then provide a case study when the input data is uniformly distributed over the unit shpere, and show that lower degree spherical harmonics are easier to be learned by over-parameterized neural networks.
****

Domain Adaptive Multibranch Networks

Róger Bermúdez-Chacón,Mathieu Salzmann,Pascal Fua

We tackle unsupervised domain adaptation by accounting for the fact that different domains may need to be processed differently to arrive to a common feature representation effective for recognition. To this end, we introduce a deep learning framework where each domain undergoes a different sequence of operations, allowing some, possibly more complex, domains to go through more computations than others.
This contrasts with state-of-the-art domain adaptation techniques that force all domains to be processed with the same series of operations, even when using multi-stream architectures whose parameters are not shared.
As evidenced by our experiments, the greater flexibility of our method translates to higher accuracy. Furthermore, it allows us to handle any number of domains simultaneously.
****

Unbiased Contrastive Divergence Algorithm for Training Energy-Based Latent Variable Models

Yixuan Qiu,Lingsong Zhang,Xiao Wang

The contrastive divergence algorithm is a popular approach to training energy-based latent variable models, which has been widely used in many machine learning models such as the restricted Boltzmann machines and deep belief nets. Despite its empirical success, the contrastive divergence algorithm is also known to have biases that severely affect its convergence. In this article we propose an unbiased version of the contrastive divergence algorithm that completely removes its bias in stochastic gradient methods, based on recent advances on unbiased Markov chain Monte Carlo methods. Rigorous theoretical analysis is developed to justify the proposed algorithm, and numerical experiments show that it significantly improves the existing method. Our findings suggest that the unbiased contrastive divergence algorithm is a promising approach to training general energy-based latent variable models.
****

Unsupervised Distillation of Syntactic Information from Contextualized Word Representations

Shauli Ravfogel,Yanai Elazar,Jacob Goldberger,Yoav Goldberg

Contextualized word representations, such as ELMo and BERT, were shown to perform well on a various of semantic and structural (syntactic) task. In this work, we tackle the task of unsupervised disentanglement between semantics and structure in neural language representations: we aim to learn a transformation of the contextualized vectors, that discards the lexical semantics, but keeps the structural information. To this end, we automatically generate groups of sentences which are structurally similar but semantically different, and use metric-learning approach to learn a transformation that emphasizes the structural component that is encoded in the vectors. We demonstrate that our transformation clusters vectors in space by structural properties, rather than by lexical semantics. Finally, we demonstrate the utility of our distilled representations by showing that they outperform the original contextualized representations in few-shot parsing set

ting.
****
Optimal Unsupervised Domain Translation
Emmanuel de Bézenac,Ibrahim Ayed,Patrick Gallinari
Unsupervised Domain Translation~(UDT) consists in finding meaningful correspondences between two domains, without access to explicit pairings between them. Following the seminal work of \textit{CycleGAN}, many variants and extensions of this model have been applied successfully to a wide range of applications. However, these methods remain poorly understood, and lack convincing theoretical guarantees. In this work, we define UDT in a rigorous, non-ambiguous manner, explore the implicit biases present in the approach and demonstrate the limits of theses approaches. Specifically, we show that mappings produced by these methods are biased towards \textit{low energy} transformations, leading us to cast UDT into an Optimal Transport~(OT) framework by making this implicit bias explicit. This not only allows us to provide theoretical guarantees for existing methods, but also to solve UDT problems where previous methods fail. Finally, making the link between the dynamic formulation of OT and CycleGAN, we propose a simple approach to solve UDT, and illustrate its properties in two distinct settings.
****
Biologically Plausible Neural Networks via Evolutionary Dynamics and Dopaminergic Plasticity
Sruthi Gorantla,Anand Louis,Christos H. Papadimitriou,Santosh Vempala,Naganand Yadati
Artificial neural networks (ANNs) lack in biological plausibility, chiefly because backpropagation requires a variant of plasticity (precise changes of the synaptic weights informed by neural events that occur downstream in the neural circuit) that is profoundly incompatible with the current understanding of the animal brain. Here we propose that backpropagation can happen in evolutionary time, instead of lifetime, in what we call neural net evolution (NNE). In NNE the weights of the links of the neural net are sparse linear functions of the animal's genes, where each gene has two alleles, 0 and 1. In each generation, a population is generated at random based on current allele frequencies, and it is tested in the learning task. The relative performance of the two alleles of each gene over the whole population is determined, and the allele frequencies are updated via the standard population genetics equations for the weak selection regime. We prove that, under assumptions, NNE succeeds in learning simple labeling functions with high probability, and with polynomially many generations and individuals per generation. We test the NNE concept, with only one hidden layer, on MNIST with encouraging results. Finally, we explore a further version of biologically plausible ANNs inspired by the recent discovery in animals of dopaminergic plasticity: the increase of the strength of a synapse that fired if dopamine was released soon after the firing.
****
Continual Learning using the SHDL Framework with Skewed Replay Distributions
Amarjot Singh,Jay McClelland
Human and animals continuously acquire, adapt as well as transfer knowledge throughout their lifespan. The ability to learn continuously is crucial for the effective functioning of agents interacting with the real world and processing continuous streams of information. Continuous learning has been a long-standing challenge for neural networks as the repeated acquisition of information from non-uniform data distributions generally lead to catastrophic forgetting or interference. This work proposes a modular architecture capable of continuous acquisition of tasks while averting catastrophic forgetting. Specifically, our contributions are: (i) Efficient Architecture: a modular architecture emulating the visual cortex that can learn meaningful representations with limited labelled examples, (ii) Knowledge Retention: retention of learned knowledge via limited replay of past experiences, (iii) Forward Transfer: efficient and relatively faster learning on new tasks, and (iv) Naturally Skewed Distributions: The learning in the above-mentioned claims is performed on non-uniform data distributions which better represent the natural statistics of our ongoing experience. Several experiments

that substantiate the above-mentioned claims are demonstrated on the CIFAR-100 dataset.

****

Differentiable Reasoning over a Virtual Knowledge Base

Bhuwan Dhingra,Manzil Zaheer,Vidhisha Balachandran,Graham Neubig,Ruslan Salakhutdinov,William W. Cohen

We consider the task of answering complex multi-hop questions using a corpus as a virtual knowledge base (KB). In particular, we describe a neural module, DrKIT, that traverses textual data like a KB, softly following paths of relations between mentions of entities in the corpus. At each step the module uses a combination of sparse-matrix TFIDF indices and a maximum inner product search (MIPS) on a special index of contextual representations of the mentions. This module is differentiable, so the full system can be trained end-to-end using gradient based methods, starting from natural language inputs. We also describe a pretraining scheme for the contextual representation encoder by generating hard negative examples using existing knowledge bases. We show that DrKIT improves accuracy by 9 points on 3-hop questions in the MetaQA dataset, cutting the gap between text-based and KB-based state-of-the-art by 70%. On HotpotQA, DrKIT leads to a 10% improvement over a BERT-based re-ranking approach to retrieving the relevant passages required to answer a question. DrKIT is also very efficient, processing up to 10-100x more queries per second than existing multi-hop systems.

****

Making Sense of Reinforcement Learning and Probabilistic Inference

Brendan O'Donoghue,Ian Osband,Catalin Ionescu

Reinforcement learning (RL) combines a control problem with statistical estimation: The system dynamics are not known to the agent, but can be learned through experience. A recent line of research casts ‘RL as inference’ and suggests a particular framework to generalize the RL problem as probabilistic inference. Our paper surfaces a key shortcoming in that approach, and clarifies the sense in which RL can be coherently cast as an inference problem. In particular, an RL agent must consider the effects of its actions upon future rewards and observations: The exploration-exploitation tradeoff. In all but the most simple settings, the resulting inference is computationally intractable so that practical RL algorithms must resort to approximation. We demonstrate that the popular ‘RL as inference’ approximation can perform poorly in even very basic problems. However, we show that with a small modification the framework does yield algorithms that can provably perform well, and we show that the resulting algorithm is equivalent to the recently proposed K-learning, which we further connect with Thompson sampling.

****

Negative Sampling in Variational Autoencoders

Adrián Csiszárik,Beatrix Benk■,Dániel Varga

We propose negative sampling as an approach to improve the notoriously bad out-of-distribution likelihood estimates of Variational Autoencoder models. Our model pushes latent images of negative samples away from the prior. When the source of negative samples is an auxiliary dataset, such a model can vastly improve on baselines when evaluated on OOD detection tasks. Perhaps more surprisingly, we present a fully unsupervised variant that can also significantly improve detection performance: using the output of the generator as a source of negative samples results in a fully unsupervised model that can be interpreted as adversarially trained.

****

Improved Training of Certifiably Robust Models

Chen Zhu,Renkun Ni,Ping-yeh Chiang,Hengduo Li,Furong Huang,Tom Goldstein

Convex relaxations are effective for training and certifying neural networks against norm-bounded adversarial attacks, but they leave a large gap between certifiable and empirical (PGD) robustness. In principle, relaxation can provide tight bounds if the convex relaxation solution is feasible for the original non-relaxed problem. Therefore, we propose two regularizers that can be used to train neu

ral networks that yield convex relaxations with tighter bounds. In all of our ex
periments, the proposed regularizations result in tighter certification bounds t
han non-regularized baselines.
****
Unsupervised Generative 3D Shape Learning from Natural Images
Attila Szabo,Givi Meishvili,Paolo Favaro
In this paper we present, to the best of our knowledge, the first method to lear
n a generative model of 3D shapes from natural images in a fully unsupervised wa
y. For example, we do not use any ground truth 3D or 2D annotations, stereo vide
o, and ego-motion during the training. Our approach follows the general strategy
 of Generative Adversarial Networks, where an image generator network learns to
create image samples that are realistic enough to fool a discriminator network i
nto believing that they are natural images. In contrast, in our approach the ima
ge gen- eration is split into 2 stages. In the first stage a generator network o
utputs 3D ob- jects. In the second, a differentiable renderer produces an image
of the 3D object from a random viewpoint. The key observation is that a realisti
c 3D object should yield a realistic rendering from any plausible viewpoint. Thu
s, by randomizing the choice of the viewpoint our proposed training forces the g
enerator network to learn an interpretable 3D representation disentangled from t
he viewpoint. In this work, a 3D representation consists of a triangle mesh and
a texture map that is used to color the triangle surface by using the UV-mapping
 technique. We provide analysis of our learning approach, expose its ambiguities
 and show how to over- come them. Experimentally, we demonstrate that our method
 can learn realistic 3D shapes of faces by using only the natural images of the
FFHQ dataset.
****
Diagnosing the Environment Bias in Vision-and-Language Navigation
Yubo Zhang,Hao Tan,Mohit Bansal
Vision-and-Language Navigation (VLN) requires an agent to follow natural-languag
e instructions, explore the given environments, and reach the desired target loc
ations. These step-by-step navigational instructions are extremely useful in nav
igating new environments which the agent does not know about previously. Most re
cent works that study VLN observe a significant performance drop when tested on
unseen environments (i.e., environments not used in training), indicating that t
he neural agent models are highly biased towards training environments. Although
 this issue is considered as one of major challenges in VLN research, it is stil
l under-studied and needs a clearer explanation. In this work, we design novel d
iagnosis experiments via environment re-splitting and feature replacement, looki
ng into possible reasons of this environment bias. We observe that neither the l
anguage nor the underlying navigational graph, but the low-level visual appearan
ce conveyed by ResNet features directly affects the agent model and contributes
to this environment bias in results. According to this observation, we explore s
everal kinds of semantic representations which contain less low-level visual inf
ormation, hence the agent learned with these features could be better generalize
d to unseen testing environments. Without modifying the baseline agent model and
 its training method, our explored semantic features significantly decrease the
performance gap between seen and unseen on multiple datasets (i.e., 8.6% to 0.2%
 on R2R, 23.9% to 0.1% on R4R, and 3.74 to 0.17 on CVDN) and achieve competitive
 unseen results to previous state-of-the-art models.
****
Learning Mahalanobis Metric Spaces via Geometric Approximation Algorithms
Diego Ihara,Neshat Mohammadi,Anastasios Sidiropoulos
Learning Mahalanobis metric spaces is an important problem that has found numero
us applications. Several algorithms have been designed for this problem, includi
ng Information Theoretic Metric Learning (ITML) [Davis et al. 2007] and Large Ma
rgin Nearest Neighbor (LMNN) classification [Weinberger and Saul 2009]. We cons
ider a formulation of Mahalanobis metric learning as an optimization problem,whe
re the objective is to minimize the number of violated similarity/dissimilarity
constraints. We show that for any fixed ambient dimension, there exists a fully
 polynomial time approximation scheme (FPTAS) with nearly-linear running time.Th

is result is obtained using tools from the theory of linear programming in low d imensions. We also discuss improvements of the algorithm in practice, and presen t experimental results on synthetic and real-world data sets. Our algorithm is f ully parallelizable and performs favorably in the presence of adversarial noise.
****
Laconic Image Classification: Human vs. Machine Performance
Javier Carrasco,Aidan Hogan,Jorge Pérez
We propose laconic classification as a novel way to understand and compare the p erformance of diverse image classifiers. The goal in this setting is to minimise the amount of information (aka. entropy) required in individual test images to maintain correct classification. Given a classifier and a test image, we compute an approximate minimal-entropy positive image for which the classifier provides a correct classification, becoming incorrect upon any further reduction. The no tion of entropy offers a unifying metric that allows to combine and compare the effects of various types of reductions (e.g., crop, colour reduction, resolution reduction) on classification performance, in turn generalising similar methods explored in previous works. Proposing two complementary frameworks for computing the minimal-entropy positive images of both human and machine classifiers, in e xperiments over the ILSVRC test-set, we find that machine classifiers are more s ensitive entropy-wise to reduced resolution (versus cropping or reduced colour f or machines, as well as reduced resolution for humans), supporting recent result s suggesting a texture bias in the ILSVRC-trained models used. We also find, in the evaluated setting, that humans classify the minimal-entropy positive images of machine models with higher precision than machines classify those of humans.
****
Prediction Poisoning: Towards Defenses Against DNN Model Stealing Attacks
Tribhuvanesh Orekondy,Bernt Schiele,Mario Fritz
High-performance Deep Neural Networks (DNNs) are increasingly deployed in many r eal-world applications e.g., cloud prediction APIs. Recent advances in model fun ctionality stealing attacks via black-box access (i.e., inputs in, predictions o ut) threaten the business model of such applications, which require a lot of tim e, money, and effort to develop. Existing defenses take a passive role against s tealing attacks, such as by truncating predicted information. We find such passi ve defenses ineffective against DNN stealing attacks. In this paper, we propose the first defense which actively perturbs predictions targeted at poisoning the training objective of the attacker. We find our defense effective across a wide range of challenging datasets and DNN model stealing attacks, and additionally o utperforms existing defenses. Our defense is the first that can withstand highly accurate model stealing attacks for tens of thousands of queries, amplifying th e attacker's error rate up to a factor of 85$\times$ with minimal impact on the utility for benign users.
****
Reinforcement Learning with Structured Hierarchical Grammar Representations of A ctions
Petros Christodoulou,Robert Lange,Ali Shafti,A. Aldo Faisal
From a young age humans learn to use grammatical principles to hierarchically co mbine words into sentences. Action grammars is the parallel idea; that there is an underlying set of rules (a "grammar") that govern how we hierarchically combi ne actions to form new, more complex actions. We introduce the Action Grammar Re inforcement Learning (AG-RL) framework which leverages the concept of action gra mmars to consistently improve the sample efficiency of Reinforcement Learning ag ents. AG-RL works by using a grammar inference algorithm to infer the “action gr ammar" of an agent midway through training, leading to a higher-level action rep resentation. The agent's action space is then augmented with macro-actions ident ified by the grammar. We apply this framework to Double Deep Q-Learning (AG-DDQN) and a discrete action version of Soft Actor-Critic (AG-SAC) and find that it i mproves performance in 8 out of 8 tested Atari games (median +31%, max +668%) an d 19 out of 20 tested Atari games (median +96%, maximum +3,756%) respectively wi thout substantive hyperparameter tuning. We also show that AG-SAC beats the mode l-free state-of-the-art for sample efficiency in 17 out of the 20 tested Atari g

ames (median +62%, maximum +13,140%), again without substantive hyperparameter t
uning.
****

The Usual Suspects? Reassessing Blame for VAE Posterior Collapse
Bin Dai,Ziyu Wang,David Wipf
In narrow asymptotic settings Gaussian VAE models of continuous data have been s
hown to possess global optima aligned with ground-truth distributions. Even so,
it is well known that poor solutions whereby the latent posterior collapses to
an uninformative prior are sometimes obtained in practice. However, contrary to
conventional wisdom that largely assigns blame for this phenomena on the undue
influence of KL-divergence regularization, we will argue that posterior collapse
is, at least in part, a direct consequence of bad local minima inherent to the
loss surface of deep autoencoder networks. In particular, we prove that even sm
all nonlinear perturbations of affine VAE decoder models can produce such minima
, and in deeper models, analogous minima can force the VAE to behave like an agg
ressive truncation operator, provably discarding information along all latent di
mensions in certain circumstances. Regardless, the underlying message here is n
ot meant to undercut valuable existing explanations of posterior collapse, but r
ather, to refine the discussion and elucidate alternative risk factors that may
have been previously underappreciated.
****

Dynamical System Embedding for Efficient Intrinsically Motivated Artificial Agen
ts
Ruihan Zhao,Stas Tiomkin,Pieter Abbeel
Mutual Information between agent Actions and environment States (MIAS) quantifie
s the influence of agent on its environment. Recently, it was found that intrins
ic motivation in artificial agents emerges from the maximization of MIAS.
For example, empowerment is an information-theoretic approach to intrinsic motiv
ation, which has been shown to solve a broad range of standard RL benchmark prob
lems. The estimation of empowerment for arbitrary dynamics is a challenging prob
lem because it relies on the estimation of MIAS. Existing approaches rely on sam
pling, which have formal limitations, requiring exponentially many samples. In t
his work, we develop a novel approach for the estimation of empowerment in unkno
wn arbitrary dynamics from visual stimulus only, without sampling for the estima
tion of MIAS. The core idea is to represent the relation between action sequence
s and future states by a stochastic dynamical system in latent space, which admi
ts an efficient estimation of MIAS by the ``Water-Filling" algorithm from inform
ation theory. We construct this embedding with deep neural networks trained on a
novel objective function and demonstrate our approach by numerical simulations
of non-linear continuous-time dynamical systems. We show that the designed embed
ding preserves information-theoretic properties of the original dynamics, and en
ables us to solve the standard AI benchmark problems.
****

SCALOR: Generative World Models with Scalable Object Representations
Jindong Jiang*,Sepehr Janghorbani*,Gerard De Melo,Sungjin Ahn
Scalability in terms of object density in a scene is a primary challenge in unsu
pervised sequential object-oriented representation learning. Most of the previou
s models have been shown to work only on scenes with a few objects. In this pape
r, we propose SCALOR, a probabilistic generative world model for learning SCALab
le Object-oriented Representation of a video. With the proposed spatially parall
el attention and proposal-rejection mechanisms, SCALOR can deal with orders of m
agnitude larger numbers of objects compared to the previous state-of-the-art mod
els. Additionally, we introduce a background module that allows SCALOR to model
complex dynamic backgrounds as well as many foreground objects in the scene. We
demonstrate that SCALOR can deal with crowded scenes containing up to a hundred
objects while jointly modeling complex dynamic backgrounds. Importantly, SCALOR
is the ■rst unsupervised object representation model shown to work for natural s
cenes containing several tens of moving objects.
****

Evaluations and Methods for Explanation through Robustness Analysis

Cheng-Yu Hsieh,Chih-Kuan Yeh,Xuanqing Liu,Pradeep Ravikumar,Seungyeon Kim,Sanjiv Kumar,Cho-Jui Hsieh
Among multiple ways of interpreting a machine learning model, measuring the importance of a set of features tied to a prediction is probably one of the most intuitive way to explain a model. In this paper, we establish the link between a set of features to a prediction with a new evaluation criteria, robustness analysis, which measures the minimum tolerance of adversarial perturbation. By measuring the tolerance level for an adversarial attack, we can extract a set of features that provides most robust support for a current prediction, and also can extract a set of features that contrasts the current prediction to a target class by setting a targeted adversarial attack. By applying this methodology to various prediction tasks across multiple domains, we observed the derived explanations are indeed capturing the significant feature set qualitatively and quantitatively.
****

Attributed Graph Learning with 2-D Graph Convolution
Qimai Li,Xiaotong Zhang,Han Liu,Xiao-Ming Wu
Graph convolutional neural networks have demonstrated promising performance in attributed graph learning, thanks to the use of graph convolution that effectively combines graph structures and node features for learning node representations. However, one intrinsic limitation of the commonly adopted 1-D graph convolution is that it only exploits graph connectivity for feature smoothing, which may lead to inferior performance on sparse and noisy real-world attributed networks. To address this problem, we propose to explore relational information among node attributes to complement node relations for representation learning. In particular, we propose to use 2-D graph convolution to jointly model the two kinds of relations and develop a computationally efficient dimensionwise separable 2-D graph convolution (DSGC). Theoretically, we show that DSGC can reduce intra-class variance of node features on both the node dimension and the attribute dimension to facilitate learning. Empirically, we demonstrate that by incorporating attribute relations, DSGC achieves significant performance gain over state-of-the-art methods on node classification and clustering on several real-world attributed networks.

****

Stochastic Neural Physics Predictor
Piotr Tatarczyk,Damian Mrowca,Li Fei-Fei,Daniel L. K. Yamins,Nils Thuerey
Recently, neural-network based forward dynamics models have been proposed that attempt to learn the dynamics of physical systems in a deterministic way. While near-term motion can be predicted accurately, long-term predictions suffer from accumulating input and prediction errors which can lead to plausible but different trajectories that diverge from the ground truth. A system that predicts distributions of the future physical states for long time horizons based on its uncertainty is thus a promising solution. In this work, we introduce a novel robust Monte Carlo sampling based graph-convolutional dropout method that allows us to sample multiple plausible trajectories for an initial state given a neural-network based forward dynamics predictor. By introducing a new shape preservation loss and training our dynamics model recurrently, we stabilize long-term predictions. We show that our model’s long-term forward dynamics prediction errors on complicated physical interactions of rigid and deformable objects of various shapes are significantly lower than existing strong baselines. Lastly, we demonstrate how generating multiple trajectories with our Monte Carlo dropout method can be used to train model-free reinforcement learning agents faster and to better solutions on simple manipulation tasks.
****

Neural tangent kernels, transportation mappings, and universal approximation
Ziwei Ji,Matus Telgarsky,Ruicheng Xian
This paper establishes rates of universal approximation for the shallow neural tangent kernel (NTK): network weights are only allowed microscopic changes from random initialization, which entails that activations are mostly unchanged, and the network is nearly equivalent to its linearization. Concretely, the paper has

two main contributions: a generic scheme to approximate functions with the NTK by sampling from transport mappings between the initial weights and their desired values, and the construction of transport mappings via Fourier transforms. Regarding the first contribution, the proof scheme provides another perspective on how the NTK regime arises from rescaling: redundancy in the weights due to resampling allows individual weights to be scaled down. Regarding the second contribution, the most notable transport mapping asserts that roughly $1 / \delta^{10d}$ nodes are sufficient to approximate continuous functions, where δ depends on the continuity properties of the target function. By contrast, nearly the same proof yields a bound of $1 / \delta^{2d}$ for shallow ReLU networks; this gap suggests a tantalizing direction for future work, separating shallow ReLU networks and their linearization.

****

Learning to Move with Affordance Maps
William Qi,Ravi Teja Mullapudi,Saurabh Gupta,Deva Ramanan
The ability to autonomously explore and navigate a physical space is a fundamental requirement for virtually any mobile autonomous agent, from household robotic vacuums to autonomous vehicles. Traditional SLAM-based approaches for exploration and navigation largely focus on leveraging scene geometry, but fail to model dynamic objects (such as other agents) or semantic constraints (such as wet floors or doorways). Learning-based RL agents are an attractive alternative because they can incorporate both semantic and geometric information, but are notoriously sample inefficient, difficult to generalize to novel settings, and are difficult to interpret. In this paper, we combine the best of both worlds with a modular approach that {\em learns} a spatial representation of a scene that is trained to be effective when coupled with traditional geometric planners. Specifically, we design an agent that learns to predict a spatial affordance map that elucidates what parts of a scene are navigable through active self-supervised experience gathering. In contrast to most simulation environments that assume a static world, we evaluate our approach in the VizDoom simulator, using large-scale randomly-generated maps containing a variety of dynamic actors and hazards. We show that learned affordance maps can be used to augment traditional approaches for both exploration and navigation, providing significant improvements in performance.
****

Towards Interpreting Deep Neural Networks via Understanding Layer Behaviors
Jiezhang Cao,Jincheng Li,Xiping Hu,Peilin Zhao,Mingkui Tan
Deep neural networks (DNNs) have achieved unprecedented practical success in many applications.
However, how to interpret DNNs is still an open problem.
In particular, what do hidden layers behave is not clearly understood.
In this paper, relying on a teacher-student paradigm, we seek to understand the layer behaviors of DNNs by ``monitoring" both across-layer and single-layer distribution evolution to some target distribution in the training. Here, the ``across-layer" and ``single-layer" considers the layer behavior \emph{along the depth} and a specific layer \emph{along training epochs}, respectively.
Relying on optimal transport theory, we employ the Wasserstein distance (W-distance) to measure the divergence between the layer distribution and the target distribution.
Theoretically, we prove that i) the W-distance of across layers to the target distribution tends to decrease along the depth. ii) the W-distance of a specific layer to the target distribution tends to decrease along training iterations. iii)
However, a deep layer is not always better than a shallow layer for some samples. Moreover, our results helps to analyze the stability of layer distributions and explains why auxiliary losses helps the training of DNNs. Extensive experiments on real-world datasets justify our theoretical findings.
****

Deep Learning For Symbolic Mathematics
Guillaume Lample,François Charton

Neural networks have a reputation for being better at solving statistical or approximate problems than at performing calculations or working with symbolic data. In this paper, we show that they can be surprisingly good at more elaborated tasks in mathematics, such as symbolic integration and solving differential equations. We propose a syntax for representing these mathematical problems, and methods for generating large datasets that can be used to train sequence-to-sequence models. We achieve results that outperform commercial Computer Algebra Systems such as Matlab or Mathematica.
****
Deep Interaction Processes for Time-Evolving Graphs
xiaofu chang,jianfeng wen,xuqin liu,yanming fang,le song,yuan qi
Time-evolving graphs are ubiquitous such as online transactions on an e-commerce platform and user interactions on social networks. While neural approaches have been proposed for graph modeling, most of them focus on static graphs. In this paper we present a principled deep neural approach that models continuous time-evolving graphs at multiple time resolutions based on a temporal point process framework. To model the dependency between latent dynamic representations of each node, we define a mixture of temporal cascades in which a node's neural representation depends on not only this node's previous representations but also the previous representations of related nodes that have interacted with this node. We generalize LSTM on this temporal cascade mixture and introduce novel time gates to model time intervals between interactions. Furthermore, we introduce a selection mechanism that gives important nodes large influence in both $k-$hop subgraphs of nodes in an interaction. To capture temporal dependency at multiple time-resolutions, we stack our neural representations in several layers and fuse them based on attention. Based on the temporal point process framework, our approach can naturally handle growth (and shrinkage) of graph nodes and interactions, making it inductive. Experimental results on interaction prediction and classification tasks -- including a real-world financial application -- illustrate the effectiveness of the time gate, the selection and attention mechanisms of our approach, as well as its
superior performance over the alternative approaches.
****
Differentiable learning of numerical rules in knowledge graphs
Po-Wei Wang,Daria Stepanova,Csaba Domokos,J. Zico Kolter
Rules over a knowledge graph (KG) capture interpretable patterns in data and can be used for KG cleaning and completion. Inspired by the TensorLog differentiable logic framework, which compiles rule inference into a sequence of differentiable operations, recently a method called Neural LP has been proposed for learning the parameters as well as the structure of rules. However, it is limited with respect to the treatment of numerical features like age, weight or scientific measurements. We address this limitation by extending Neural LP to learn rules with numerical values, e.g., ”People younger than 18 typically live with their parents“. We demonstrate how dynamic programming and cumulative sum operations can be exploited to ensure efficiency of such extension. Our novel approach allows us to extract more expressive rules with aggregates, which are of higher quality and yield more accurate predictions compared to rules learned by the state-of-the-art methods, as shown by our experiments on synthetic and real-world datasets.
****
Consistency Regularization for Generative Adversarial Networks
Han Zhang,Zizhao Zhang,Augustus Odena,Honglak Lee
Generative Adversarial Networks (GANs) are known to be difficult to train, despite considerable research effort. Several regularization techniques for stabilizing training have been proposed, but they introduce non-trivial computational overheads and interact poorly with existing techniques like spectral normalization. In this work, we propose a simple, effective training stabilizer based on the notion of consistency regularization—a popular technique in the semi-supervised learning literature. In particular, we augment data passing into the GAN discriminator and penalize the sensitivity of the discriminator to these augmentations. We conduct a series of experiments to demonstrate that consistency regularizatio

n works effectively with spectral normalization and various GAN architectures, l oss functions and optimizer settings. Our method achieves the best FID scores fo r unconditional image generation compared to other regularization methods on CIF AR-10 and CelebA. Moreover, Our consistency regularized GAN (CR-GAN) improves st ate of-the-art FID scores for conditional generation from 14.73 to 11.48 on CIFA R-10 and from 8.73 to 6.66 on ImageNet-2012.
****

On Generalization Error Bounds of Noisy Gradient Methods for Non-Convex Learning
Jian Li,Xuanyuan Luo,Mingda Qiao
Generalization error (also known as the out-of-sample error) measures how well t he hypothesis learned from training data generalizes to previously unseen data. Proving tight generalization error bounds is a central question in statistical l earning theory. In this paper, we obtain generalization error bounds for learning general non-convex objectives, which has attracted significa nt attention in recent years. We develop a new framework, termed Bayes-Sta bility, for proving algorithm-dependent generalization error bounds. The new f ramework combines ideas from both the PAC-Bayesian theory and the notion of algo rithmic stability. Applying the Bayes-Stability method, we obtain new data-depe ndent generalization bounds for stochastic gradient Langevin dynamics (SGLD) and several other noisy gradient methods (e.g., with momentum, mini-batch and accel eration, Entropy-SGD). Our result recovers (and is typically tighter than) a rec ent result in Mou et al. (2018) and improves upon the results in Pensia et al. (2018). Our experiments demonstrate that our data-dependent bounds can distingui sh randomly labelled data from normal data, which provides an explanation to the intriguing phenomena observed in Zhang et al. (2017a). We also study the settin g where the total loss is the sum of a bounded loss and an additiona l`2 regular ization term. We obtain new generalization bounds for the continuous Langevin dy namic in this setting by developing a new Log-Sobolev inequality for the paramet er distribution at any time. Our new bounds are more desirable when the noise le vel of the processis not very small, and do not become vacuous even when T tends to infinity.
****

SUMO: Unbiased Estimation of Log Marginal Probability for Latent Variable Models
Yucen Luo,Alex Beatson,Mohammad Norouzi,Jun Zhu,David Duvenaud,Ryan P. Adams,Ric ky T. Q. Chen
Standard variational lower bounds used to train latent variable models produce b iased estimates of most quantities of interest. We introduce an unbiased estimat or of the log marginal likelihood and its gradients for latent variable models b ased on randomized truncation of infinite series. If parameterized by an encoder -decoder architecture, the parameters of the encoder can be optimized to minimiz e its variance of this estimator. We show that models trained using our estimato r give better test-set likelihoods than a standard importance-sampling based app roach for the same average computational cost. This estimator also allows use of latent variable models for tasks where unbiased estimators, rather than margina l likelihood lower bounds, are preferred, such as minimizing reverse KL divergen ces and estimating score functions.
****

Benefits of Overparameterization in Single-Layer Latent Variable Generative Mode ls
Rares-Darius Buhai,Andrej Risteski,Yoni Halpern,David Sontag
One of the most surprising and exciting discoveries in supervising learning was the benefit of overparameterization (i.e. training a very large model) to improv ing the optimization landscape of a problem, with minimal effect on statistical performance (i.e. generalization). In contrast, unsupervised settings have been under-explored, despite the fact that it has been observed that overparameteriza tion can be helpful as early as Dasgupta & Schulman (2007). In this paper, we pe rform an exhaustive study of different aspects of overparameterization in unsupe rvised learning via synthetic and semi-synthetic experiments. We discuss benefit s to different metrics of success (recovering the parameters of the ground-truth model, held-out log-likelihood), sensitivity to variations of the training algo

rithm, and behavior as the amount of overparameterization increases. We find tha
t, when learning using methods such as variational inference, larger models ca
n significantly increase the number of ground truth latent variables recovered.
****

Implicit competitive regularization in GANs
Florian Schaefer,Hongkai Zheng,Anima Anandkumar
Generative adversarial networks (GANs) are capable of producing high quality sam
ples, but they suffer from numerous issues such as instability and mode collapse
 during training. To combat this, we propose to model the generator and discrimi
nator as agents acting under local information, uncertainty, and awareness of th
eir opponent. By doing so we achieve stable convergence, even when the underlyin
g game has no Nash equilibria. We call this mechanism \emph{implicit competitiv
e regularization} (ICR) and show that it is present in the recently proposed \em
ph{competitive gradient descent} (CGD).
When comparing CGD to Adam using a variety of loss functions and regularizers on
 CIFAR10, CGD shows a much more consistent performance, which we attribute to IC
R.
In our experiments, we achieve the highest inception score when using the WGAN l
oss (without gradient penalty or weight clipping) together with CGD. This can be
 interpreted as minimizing a form of integral probability metric based on ICR.
****

Scale-Equivariant Steerable Networks
Ivan Sosnovik,Micha■ Szmaja,Arnold Smeulders
The effectiveness of Convolutional Neural Networks (CNNs) has been substantially
 attributed to their built-in property of translation equivariance. However, CNN
s do not have embedded mechanisms to handle other types of transformations. In t
his work, we pay attention to scale changes, which regularly appear in various t
asks due to the changing distances between the objects and the camera. First, we
 introduce the general theory for building scale-equivariant convolutional netwo
rks with steerable filters. We develop scale-convolution and generalize other co
mmon blocks to be scale-equivariant. We demonstrate the computational efficiency
 and numerical stability of the proposed method. We compare the proposed models
to the previously developed methods for scale equivariance and local scale invar
iance. We demonstrate state-of-the-art results on the MNIST-scale dataset and on
 the STL-10 dataset in the supervised learning setting.
****

DeepSphere: a graph-based spherical CNN
Michaël Defferrard,Martino Milani,Frédérick Gusset,Nathanaël Perraudin
Designing a convolution for a spherical neural network requires a delicate trade
off between efficiency and rotation equivariance. DeepSphere, a method based on
a graph representation of the discretized sphere, strikes a controllable balance
 between these two desiderata. This contribution is twofold. First, we study bot
h theoretically and empirically how equivariance is affected by the underlying g
raph with respect to the number of pixels and neighbors. Second, we evaluate Dee
pSphere on relevant problems. Experiments show state-of-the-art performance and
demonstrates the efficiency and flexibility of this formulation. Perhaps surpris
ingly, comparison with previous work suggests that anisotropic filters might be
an unnecessary price to pay. Our code is available at https://github.com/deepsph
ere.
****

Improved Training Techniques for Online Neural Machine Translation
Maha Elbayad,Laurent Besacier,Jakob Verbeek
Neural sequence-to-sequence models are at the basis of state-of-the-art solution
s for sequential prediction problems such as machine translation and speech reco
gnition. The models typically assume that the entire input is available when sta
rting target generation. In some applications, however, it is desirable to start
 the decoding process before the entire input is available, e.g. to reduce the l
atency in automatic speech recognition. We consider state-of-the-art wait-k deco
ders, that first read k tokens from the source and then alternate between readin
g tokens from the input and writing to the output. We investigate the sensitivit

y of such models to the value of k that is used during training and when deploying the model, and the effect of updating the hidden states in transformer models as new source tokens are read. We experiment with German-English translation on the IWSLT14 dataset and the larger WMT15 dataset. Our results significantly improve over earlier state-of-the-art results for German-English translation on the WMT15 dataset across different latency levels.
****

GRASPEL: GRAPH SPECTRAL LEARNING AT SCALE
Yongyu Wang,Zhiqiang Zhao,Zhuo Feng
Learning meaningful graphs from data plays important roles in many data mining and machine learning tasks, such as data representation and analysis, dimension reduction, data clustering, and visualization, etc. In this work, we present a scalable spectral approach to graph learning from data. By limiting the precision matrix to be a graph Laplacian, our approach aims to estimate ultra-sparse weighted graphs and has a clear connection with the prior graphical Lasso method. By interleaving nearly-linear time spectral graph sparsification, coarsening and embedding procedures, ultra-sparse yet spectrally-stable graphs can be iteratively constructed in a highly-scalable manner. Compared with prior graph learning approaches that do not scale to large problems, our approach is highly-scalable for constructing graphs that can immediately lead to substantially improved computing efficiency and solution quality for a variety of data mining and machine learning applications, such as spectral clustering (SC), and t-Distributed Stochastic Neighbor Embedding (t-SNE).
****

Overcoming Catastrophic Forgetting via Hessian-free Curvature Estimates
Leonid Butyrev,Georgios Kontes,Christoffer Löffler,Christopher Mutschler
Learning neural networks with gradient descent over a long sequence of tasks is problematic as their fine-tuning to new tasks overwrites the network weights that are important for previous tasks. This leads to a poor performance on old tasks – a phenomenon framed as catastrophic forgetting. While early approaches use task rehearsal and growing networks that both limit the scalability of the task sequence orthogonal approaches build on regularization. Based on the Fisher information matrix (FIM) changes to parameters that are relevant to old tasks are penalized, which forces the task to be mapped into the available remaining capacity of the network. This requires to calculate the Hessian around a mode, which makes learning tractable. In this paper, we introduce Hessian-free curvature estimates as an alternative method to actually calculating the Hessian. In contrast to previous work, we exploit the fact that most regions in the loss surface are flat and hence only calculate a Hessian-vector-product around the surface that is relevant for the current task. Our experiments show that on a variety of well-known task sequences we either significantly outperform or are en par with previous work.
****

Score and Lyrics-Free Singing Voice Generation
Jen-Yu Liu,Yu-Hua Chen,Yin-Cheng Yeh,Yi-Hsuan Yang
Generative models for singing voice have been mostly concerned with the task of "singing voice synthesis," i.e., to produce singing voice waveforms given musical scores and text lyrics. In this work, we explore a novel yet challenging alternative: singing voice generation without pre-assigned scores and lyrics, in both training and inference time. In particular, we experiment with three different schemes: 1) free singer, where the model generates singing voices without taking any conditions; 2) accompanied singer, where the model generates singing voices over a waveform of instrumental music; and 3) solo singer, where the model improvises a chord sequence first and then uses that to generate voices. We outline the associated challenges and propose a pipeline to tackle these new tasks. This involves the development of source separation and transcription models for data preparation, adversarial networks for audio generation, and customized metrics for evaluation.
****

Neural Video Encoding

Abel Brown,Robert DiPietro
Deep neural networks have had unprecedented success in computer vision, natural language processing, and speech largely due to the ability to search for suitable task algorithms via differentiable programming. In this paper, we borrow ideas from Kolmogorov complexity theory and normalizing flows to explore the possibilities of finding arbitrary algorithms that represent data. In particular, algorithms which encode sequences of video image frames. Ultimately, we demonstrate neural video encoded using convolutional neural networks to transform autoregressive noise processes and show that this method has surprising cryptographic analogs for information security.
****

Classification-Based Anomaly Detection for General Data
Liron Bergman,Yedid Hoshen
Anomaly detection, finding patterns that substantially deviate from those seen previously, is one of the fundamental problems of artificial intelligence. Recently, classification-based methods were shown to achieve superior results on this task. In this work, we present a unifying view and propose an open-set method, GOAD, to relax current generalization assumptions. Furthermore, we extend the applicability of transformation-based methods to non-image data using random affine transformations. Our method is shown to obtain state-of-the-art accuracy and is applicable to broad data types. The strong performance of our method is extensively validated on multiple datasets from different domains.
****

Distributed Training Across the World
Ligeng Zhu,Yao Lu,Yujun Lin,Song Han
Traditional synchronous distributed training is performed inside a cluster, since it requires high bandwidth and low latency network (e.g. 25Gb Ethernet or Infini-band). However, in many application scenarios, training data are often distributed across many geographic locations, where physical distance is long and latency is high. Traditional synchronous distributed training cannot scale well under such limited network conditions. In this work, we aim to scale distributed learning un-der high-latency network. To achieve this, we propose delayed and temporally sparse (DTS) update that enables synchronous training to tolerate extreme network conditions without compromising accuracy. We benchmark our algorithms on servers deployed across three continents in the world: London (Europe), Tokyo (Asia), Oregon (North America) and Ohio (North America). Under such challenging settings, DTS achieves90×speedup over traditional methods without loss of accuracy on ImageNet.
****

Unrestricted Adversarial Examples via Semantic Manipulation
Anand Bhattad,Min Jin Chong,Kaizhao Liang,Bo Li,D. A. Forsyth
Machine learning models, especially deep neural networks (DNNs), have been shown to be vulnerable against adversarial examples which are carefully crafted samples with a small magnitude of the perturbation. Such adversarial perturbations are usually restricted by bounding their $\mathcal{L}_p$ norm such that they are imperceptible, and thus many current defenses can exploit this property to reduce their adversarial impact. In this paper, we instead introduce "unrestricted" perturbations that manipulate semantically meaningful image-based visual descriptors - color and texture - in order to generate effective and photorealistic adversarial examples. We show that these semantically aware perturbations are effective against JPEG compression, feature squeezing and adversarially trained model. We also show that the proposed methods can effectively be applied to both image classification and image captioning tasks on complex datasets such as ImageNet and MSCOCO. In addition, we conduct comprehensive user studies to show that our generated semantic adversarial examples are photorealistic to humans despite large magnitude perturbations when compared to other attacks.
****

Stochastic Latent Actor-Critic: Deep Reinforcement Learning with a Latent Variable Model
Alex X. Lee,Anusha Nagabandi,Pieter Abbeel,Sergey Levine

Deep reinforcement learning (RL) algorithms can use high-capacity deep networks to learn directly from image observations. However, these kinds of observation s paces present a number of challenges in practice, since the policy must now sol ve two problems: a representation learning problem, and a task learning problem. In this paper, we aim to explicitly learn representations that can accelerate r einforcement learning from images. We propose the stochastic latent actor-critic (SLAC) algorithm: a sample-efficient and high-performing RL algorithm for learn ing policies for complex continuous control tasks directly from high-dimensional image inputs. SLAC learns a compact latent representation space using a stochas tic sequential latent variable model, and then learns a critic model within this latent space. By learning a critic within a compact state space, SLAC can learn much more efficiently than standard RL methods. The proposed model improves per formance substantially over alternative representations as well, such as variati onal autoencoders. In fact, our experimental evaluation demonstrates that the sa mple efficiency of our resulting method is comparable to that of model-based RL methods that directly use a similar type of model for control. Furthermore, our method outperforms both model-free and model-based alternatives in terms of fina l performance and sample efficiency, on a range of difficult image-based control tasks. Our code and videos of our results are available at our website.
****

Closed loop deep Bayesian inversion: Uncertainty driven acquisition for fast MR I

Thomas Sanchez,Igor Krawczuk,Zhaodong Sun,Volkan Cevher

This work proposes a closed-loop, uncertainty-driven adaptive sampling frame- wo rk (CLUDAS) for accelerating magnetic resonance imaging (MRI) via deep Bayesian inversion. By closed-loop, we mean that our samples adapt in real- time to the i ncoming data. To our knowledge, we demonstrate the first genera- tive adversaria l network (GAN) based framework for posterior estimation over a continuum sampli ng rates of an inverse problem. We use this estimator to drive the sampling for accelerated MRI. Our numerical evidence demonstrates that the variance estimate strongly correlates with the expected MSE improvement for dif- ferent accelerati on rates even with few posterior samples. Moreover, the resulting masks bring im provements to the state-of-the-art fixed and active mask designing approaches ac ross MSE, posterior variance and SSIM on real undersampled MRI scans.
****

OBJECT-ORIENTED REPRESENTATION OF 3D SCENES

Chang Chen,Sungjin Ahn

In this paper, we propose a generative model, called ROOTS (Representation of Ob ject-Oriented Three-dimension Scenes), for unsupervised object-wise 3D-scene dec omposition and and rendering. For 3D scene modeling, ROOTS bases on the Generati ve Query Networks (GQN) framework, but unlike GQN, provides object-oriented repr esentation decomposition. The inferred object-representation of ROOTS is 3D in t he sense that it is viewpoint invariant as the full scene representation of GQN is so. ROOTS also provides hierarchical object-oriented representation: at 3D gl obal-scene level and at 2D local-image level. We achieve this without performan ce degradation. In experiments on datasets of 3D rooms with multiple objects, we demonstrate the above properties by focusing on its abilities for disentangleme nt, compositionality, and generalization in comparison to GQN.
****

Discriminative Particle Filter Reinforcement Learning for Complex Partial observ ations

Xiao Ma,Peter Karkus,David Hsu,Wee Sun Lee,Nan Ye

Deep reinforcement learning is successful in decision making for sophisticated g ames, such as Atari, Go, etc.

However, real-world decision making often requires reasoning with partial inform ation extracted from complex visual observations. This paper presents Discrimin ative Particle Filter Reinforcement Learning (DPFRL), a new reinforcement learni ng framework for complex partial observations. DPFRL encodes a differentiable pa rticle filter in the neural network policy for explicit reasoning with partial o bservations over time. The particle filter maintains a belief using learned disc

riminative update, which is trained end-to-end for decision making. We show that using the discriminative update instead of standard generative models results in significantly improved performance, especially for tasks with complex visual observations, because they circumvent the difficulty of modeling complex observations that are irrelevant to decision making.
In addition, to extract features from the particle belief, we propose a new type of belief feature based on the moment generating function. DPFRL outperforms state-of-the-art POMDP RL models in Flickering Atari Games, an existing POMDP RL benchmark, and in Natural Flickering Atari Games, a new, more challenging POMDP RL benchmark introduced in this paper. Further, DPFRL performs well for visual navigation with real-world data in the Habitat environment.
****
Learning to Group: A Bottom-Up Framework for 3D Part Discovery in Unseen Categories
Tiange Luo,Kaichun Mo,Zhiao Huang,Jiarui Xu,Siyu Hu,Liwei Wang,Hao Su
We address the problem of learning to discover 3D parts for objects in unseen categories. Being able to learn the geometry prior of parts and transfer this prior to unseen categories pose fundamental challenges on data-driven shape segmentation approaches. Formulated as a contextual bandit problem, we propose a learning-based iterative grouping framework which learns a grouping policy to progressively merge small part proposals into bigger ones in a bottom-up fashion. At the core of our approach is to restrict the local context for extracting part-level features, which encourages the generalizability to novel categories. On a recently proposed large-scale fine-grained 3D part dataset, PartNet, we demonstrate that our method can transfer knowledge of parts learned from 3 training categories to 21 unseen testing categories without seeing any annotated samples. Quantitative comparisons against four strong shape segmentation baselines show that we achieve the state-of-the-art performance.
****
State Alignment-based Imitation Learning
Fangchen Liu,Zhan Ling,Tongzhou Mu,Hao Su
Consider an imitation learning problem that the imitator and the expert have different dynamics models. Most of existing imitation learning methods fail because they focus on the imitation of actions. We propose a novel state alignment-based imitation learning method to train the imitator by following the state sequences in the expert demonstrations as much as possible. The alignment of states comes from both local and global perspectives. We combine them into a reinforcement learning framework by a regularized policy update objective. We show the superiority of our method on standard imitation learning settings as well as the challenging settings in which the expert and the imitator have different dynamics models.
****
Reweighted Proximal Pruning for Large-Scale Language Representation
Fu-Ming Guo,Sijia Liu,Finlay S. Mungall,Xue Lin,Yanzhi Wang
Recently, pre-trained language representation flourishes as the mainstay of the natural language understanding community, e.g., BERT. These pre-trained language representations can create state-of-the-art results on a wide range of downstream tasks. Along with continuous significant performance improvement, the size and complexity of these pre-trained neural models continue to increase rapidly. Is it possible to compress these large-scale language representation models? How will the pruned language representation affect the downstream multi-task transfer learning objectives? In this paper, we propose Reweighted Proximal Pruning (RPP), a new pruning method specifically designed for a large-scale language representation model. Through experiments on SQuAD and the GLUE benchmark suite, we show that proximal pruned BERT keeps high accuracy for both the pre-training task and the downstream multiple fine-tuning tasks at high prune ratio. RPP provides a new perspective to help us analyze what large-scale language representation might learn. Additionally, RPP makes it possible to deploy a large state-of-the-art language representation model such as BERT on a series of distinct devices (e.g., online servers, mobile phones, and edge devices).

****
Neural Arithmetic Units
Andreas Madsen,Alexander Rosenberg Johansen
Neural networks can approximate complex functions, but they struggle to perform exact arithmetic operations over real numbers. The lack of inductive bias for arithmetic operations leaves neural networks without the underlying logic necessary to extrapolate on tasks such as addition, subtraction, and multiplication. We present two new neural network components: the Neural Addition Unit (NAU), which can learn exact addition and subtraction; and the Neural Multiplication Unit (NMU) that can multiply subsets of a vector. The NMU is, to our knowledge, the first arithmetic neural network component that can learn to multiply elements from a vector, when the hidden size is large. The two new components draw inspiration from a theoretical analysis of recently proposed arithmetic components. We find that careful initialization, restricting parameter space, and regularizing for sparsity is important when optimizing the NAU and NMU. Our proposed units NAU and NMU, compared with previous neural units, converge more consistently, have fewer parameters, learn faster, can converge for larger hidden sizes, obtain sparse and meaningful weights, and can extrapolate to negative and small values.
****
Lipschitz constant estimation of Neural Networks via sparse polynomial optimization
Fabian Latorre,Paul Rolland,Volkan Cevher
We introduce LiPopt, a polynomial optimization framework for computing increasingly tighter upper bound on the Lipschitz constant of neural networks. The underlying optimization problems boil down to either linear (LP) or semidefinite (SDP) programming. We show how to use the sparse connectivity of a network, to significantly reduce the complexity of computation. This is specially useful for convolutional as well as pruned neural networks. We conduct experiments on networks with random weights as well as networks trained on MNIST, showing that in the particular case of the ℓ_∞-Lipschitz constant, our approach yields superior estimates as compared to other baselines available in the literature.

****
Random Bias Initialization Improving Binary Neural Network Training
Xinlin Li,Vahid Partovi Nia
Edge intelligence especially binary neural network (BNN) has attracted considerable attention of the artificial intelligence community recently. BNNs significantly reduce the computational cost, model size, and memory footprint. However, there is still a performance gap between the successful full-precision neural network with ReLU activation and BNNs. We argue that the accuracy drop of BNNs is due to their geometry.
We analyze the behaviour of the full-precision neural network with ReLU activation and compare it with its binarized counterpart. This comparison suggests random bias initialization as a remedy to activation saturation in full-precision networks and leads us towards an improved BNN training. Our numerical experiments confirm our geometric intuition.
****
Meta-RCNN: Meta Learning for Few-Shot Object Detection
Xiongwei Wu,Doyen Sahoo,Steven C. H. Hoi
Despite significant advances in object detection in recent years, training effective detectors in a small data regime remains an open challenge. Labelling training data for object detection is extremely expensive, and there is a need to develop techniques that can generalize well from small amounts of labelled data. We investigate this problem of few-shot object detection, where a detector has access to only limited amounts of annotated data. Based on the recently evolving meta-learning principle, we propose a novel meta-learning framework for object detection named ``Meta-RCNN", which learns the ability to perform few-shot detection via meta-learning. Specifically, Meta-RCNN learns an object detector in an episodic learning paradigm on the (meta) training data. This learning scheme helps acquire a prior which enables Meta-RCNN to do few-shot detection on novel tasks.

Built on top of the Faster RCNN model, in Meta-RCNN, both the Region Proposal Network (RPN) and the object classification branch are meta-learned. The meta-trained RPN learns to provide class-specific proposals, while the object classifier learns to do few-shot classification. The novel loss objectives and learning strategy of Meta-RCNN can be trained in an end-to-end manner. We demonstrate the effectiveness of Meta-RCNN in addressing few-shot detection on Pascal VOC dataset and achieve promising results.

****

Adversarially learned anomaly detection for time series data

Alexander Geiger,Alfredo Cuesta-Infante,Kalyan Veeramachaneni

Anomaly detection in time series data is an important topic in many domains. However, time series are known to be particular hard to analyze. Based on the recent developments in adversarially learned models, we propose a new approach for anomaly detection in time series data. We build upon the idea to use a combination of a reconstruction error and the output of a Critic network. To this end we propose a cycle-consistent GAN architecture for sequential data and a new way of measuring the reconstruction error. We then show in a detailed evaluation how the different parts of our model contribute to the final anomaly score and demonstrate how the method improves the results on several data sets. We also compare our model to other baseline anomaly detection methods to verify its performance.

****

Effect of Activation Functions on the Training of Overparametrized Neural Nets

Abhishek Panigrahi,Abhishek Shetty,Navin Goyal

It is well-known that overparametrized neural networks trained using gradient based methods quickly achieve small training error with appropriate hyperparameter settings. Recent papers have proved this statement theoretically for highly overparametrized networks under reasonable assumptions. These results either assume that the activation function is ReLU or they depend on the minimum eigenvalue of a certain Gram matrix. In the latter case, existing works only prove that this minimum eigenvalue is non-zero and do not provide quantitative bounds which require that this eigenvalue be large. Empirically, a number of alternative activation functions have been proposed which tend to perform better than ReLU at least in some settings but no clear understanding has emerged. This state of affairs underscores the importance of theoretically understanding the impact of activation functions on training. In the present paper, we provide theoretical results about the effect of activation function on the training of highly overparametrized 2-layer neural networks. A crucial property that governs the performance of an activation is whether or not it is smooth:

• For non-smooth activations such as ReLU, SELU, ELU, which are not smooth because there is a point where either the ■rst order or second order derivative is discontinuous, all eigenvalues of the associated Gram matrix are large under minimal assumptions on the data.

• For smooth activations such as tanh, swish, polynomial, which have derivatives of all orders at all points, the situation is more complex: if the subspace spanned by the data has small dimension then the minimum eigenvalue of the Gram matrix can be small leading to slow training. But if the dimension is large and the data satis■es another mild condition, then the eigenvalues are large. If we allow deep networks, then the small data dimension is not a limitation provided that the depth is suf■cient.

We discuss a number of extensions and applications of these results.

****

Multi-Precision Policy Enforced Training (MuPPET) : A precision-switching strategy for quantised fixed-point training of CNNs

Aditya Rajagopal,Diederik A. Vink,Stylianos I. Venieris,Christos-Savvas Bouganis

Large-scale convolutional neural networks (CNNs) suffer from very long training times, spanning from hours to weeks, limiting the productivity and experimentation of deep learning practitioners. As networks grow in size and complexity one approach of reducing training time is the use of low-precision data representation and computations during the training stage. However, in doing so the final accuracy suffers due to the problem of vanishing gradients. Existing

state-of-the-art methods combat this issue by means of a mixed-precision approach employing two different precision levels, FP32 (32-bit floating-point precision) and FP16/FP8 (16-/8-bit floating-point precision), leveraging the hardware support of recent GPU architectures for FP16 operations to obtaining performance gains. This work pushes the boundary of quantised training by employing a multilevel optimisation approach that utilises multiple precisions including low-precision fixed-point representations. The training strategy, named MuPPET, combines the use of multiple number representation regimes together with a precision-switching mechanism that decides at run time the transition between different precisions. Overall, the proposed strategy tailors the training process to the hardware-level capabilities of the utilised hardware architecture and yields improvements in training time and energy efficiency compared to state-of-the-art approaches. Applying MuPPET on the training of AlexNet, ResNet18 and GoogLeNet on ImageNet (ILSVRC12) and targeting an NVIDIA Turing GPU, the proposed method achieves the same accuracy as the standard full-precision training with an average training-time speedup of 1.28× across the networks.

****

Deep Spike Decoder (DSD)

Emrah Adamey,Tarin Ziyaee,Nishanth Alapati,Jun Ye

Spike-sorting is of central importance for neuroscience research. We introducea novel spike-sorting method comprising a deep autoencoder trained end-to-endwith a biophysical generative model, biophysically motivated priors, and a self-supervised loss function to training a deep autoencoder. The encoder infers the ac-tion potential event times for each source, while the decoder parameters representeach source’s spatiotemporal response waveform. We evaluate this approach inthe context of real and synthetic multi-channel surface electromyography (sEMG)data, a noisy superposition of motor unit action potentials (MUAPs). Relative toan established spike-sorting method, this autoencoder-based approach shows su-perior recovery of source waveforms and event times. Moreover, the biophysicalnature of the loss functions facilitates interpretability and hyperparameter tuning.Overall, these results demonstrate the efficacy and motivate further developmentof self-supervised spike sorting techniques.

****

Isolating Latent Structure with Cross-population Variational Autoencoders

Joe Davison,Kristen A. Severson,Soumya Ghosh

A significant body of recent work has examined variational autoencoders as a powerful approach for tasks which involve modeling the distribution of complex data such as images and text. In this work, we present a framework for modeling multiple data sets which come from differing distributions but which share some common latent structure. By incorporating architectural constraints and using a mutual information regularized form of the variational objective, our method successfully models differing data populations while explicitly encouraging the isolation of the shared and private latent factors. This enables our model to learn useful shared structure across similar tasks and to disentangle cross-population representations in a weakly supervised way. We demonstrate the utility of our method on several applications including image denoising, sub-group discovery, and continual learning.

****

Learning Compact Embedding Layers via Differentiable Product Quantization

Ting Chen,Lala Li,Yizhou Sun

Embedding layers are commonly used to map discrete symbols into continuous embedding vectors that reflect their semantic meanings. Despite their effectiveness, the number of parameters in an embedding layer increases linearly with the number of symbols and poses a critical challenge on memory and storage constraints. In this work, we propose a generic and end-to-end learnable compression framework termed differentiable product quantization (DPQ). We present two instantiations of DPQ that leverage different approximation techniques to enable differentiability in end-to-end learning. Our method can readily serve as a drop-in alternative for any existing embedding layer. Empirically, DPQ offers significant compression ratios (14-238x) at negligible or no performance cost on 10 datasets across

three different language tasks.
****
Accelerating First-Order Optimization Algorithms
Ange Tato,Roger Nkambou
Several stochastic optimization algorithms are currently available. In most cases, selecting the best optimizer for a given problem is not an easy task. Therefore, instead of looking for yet another 'absolute' best optimizer, accelerating existing ones according to the context might prove more effective. This paper presents a simple and intuitive technique to accelerate first-order optimization algorithms. When applied to first-order optimization algorithms, it converges much more quickly and achieves lower function/loss values when compared to traditional algorithms. The proposed solution modifies the update rule, based on the variation of the direction of the gradient during training. Several tests were conducted with SGD, AdaGrad, Adam and AMSGrad on three public datasets. Results clearly show that the proposed technique, has the potential to improve the performance of existing optimization algorithms.
****
Physics-Aware Flow Data Completion Using Neural Inpainting
Sebastien Foucher,Jingwei Tang,Vinicius da Costa de Azevedo,Byungsoo Kim,Markus Gross,Barbara Solenthaler
In this paper we propose a physics-aware neural network for inpainting fluid flow data. We consider that flow field data inherently follows the solution of the Navier-Stokes equations and hence our network is designed to capture physical laws. We use a DenseBlock U-Net architecture combined with a stream function formulation to inpaint missing velocity data. Our loss functions represent the relevant physical quantities velocity, velocity Jacobian, vorticity and divergence. Obstacles are treated as known priors, and each layer of the network receives the relevant information through concatenation with the previous layer's output. Our results demonstrate the network's capability for physics-aware completion tasks, and the presented ablation studies show the effectiveness of each proposed component.
****
Imagine That! Leveraging Emergent Affordances for Tool Synthesis in Reaching Tasks
Yizhe Wu,Sudhanshu Kasewa,Oliver Groth,Sasha Salter,Li Sun,Oiwi Parker Jones,Ingmar Posner
In this paper we investigate an artificial agent's ability to perform task-focused tool synthesis via imagination. Our motivation is to explore the richness of information captured by the latent space of an object-centric generative model - and how to exploit it. In particular, our approach employs activation maximisation of a task-based performance predictor to optimise the latent variable of a structured latent-space model in order to generate tool geometries appropriate for the task at hand. We evaluate our model using a novel dataset of synthetic reaching tasks inspired by the cognitive sciences and behavioural ecology. In doing so we examine the model's ability to imagine tools for increasingly complex scenario types, beyond those seen during training. Our experiments demonstrate that the synthesis process modifies emergent, task-relevant object affordances in a targeted and deliberate way: the agents often specifically modify aspects of the tools which relate to meaningful (yet implicitly learned) concepts such as a tool's length, width and configuration. Our results therefore suggest, that task relevant object affordances are implicitly encoded as directions in a structured latent space shaped by experience.
****
Provable Filter Pruning for Efficient Neural Networks
Lucas Liebenwein,Cenk Baykal,Harry Lang,Dan Feldman,Daniela Rus
We present a provable, sampling-based approach for generating compact Convolutional Neural Networks (CNNs) by identifying and removing redundant filters from an over-parameterized network. Our algorithm uses a small batch of input data points to assign a saliency score to each filter and constructs an importance sampling distribution where filters that highly affect the output are sampled with cor

respondingly high probability.
In contrast to existing filter pruning approaches, our method is simultaneously data-informed, exhibits provable guarantees on the size and performance of the pruned network, and is widely applicable to varying network architectures and data sets. Our analytical bounds bridge the notions of compressibility and importance of network structures, which gives rise to a fully-automated procedure for identifying and preserving filters in layers that are essential to the network's performance. Our experimental evaluations on popular architectures and data sets show that our algorithm consistently generates sparser and more efficient models than those constructed by existing filter pruning approaches.
****
ADAPTIVE GENERATION OF PROGRAMMING PUZZLES
Ashwin Kalyan,Oleksandr Polozov,Adam Tauman Kalai
AI today is far from being able to write complex programs. What type of problems would be best for computers to learn to program, and how should such problems be generated? To answer the first question, we suggest programming puzzles as a domain for teaching computers programming. A programming puzzle consists of a short program for a Boolean function f(x) and the goal is, given the source code, to
find an input that makes f return True. Puzzles are objective in that one can easily
test the correctness of a given solution x by seeing whether it satisfies f, unlike the
most common representations for program synthesis: given input-output pairs or an
English problem description, the correctness of a given solution is not determined
and is debatable. To address the second question of automatic puzzle generation, we suggest a GAN-like generation algorithm called “Troublemaker” which can generate puzzles targeted at any given puzzle-solver. The main innovation is that it
adapts to one or more given puzzle-solvers: rather than generating a single data set
of puzzles, Tro
****
Learning transitional skills with intrinsic motivation
Qiangxing Tian,Jinxin Liu,Donglin Wang
By maximizing an information theoretic objective, a few recent methods empower the agent to explore the environment and learn useful skills without supervision. However, when considering to use multiple consecutive skills to complete a specific task, the transition from one to another cannot guarantee the success of the process due to the evident gap between skills. In this paper, we propose to learn transitional skills (LTS) in addition to creating diverse primitive skills without a reward function. By introducing an extra latent variable for transitional skills, our LTS method discovers both primitive and transitional skills by minimizing the difference of mutual information and the similarity of skills. By considering various simulated robotic tasks, our results demonstrate the effectiveness of LTS on learning both diverse primitive skills and transitional skills, and show its superiority in smooth transition of skills over the state-of-the-art baseline DIAYN.
****
Quantifying uncertainty with GAN-based priors
Dhruv V. Patel,Assad A. Oberai
Bayesian inference is used extensively to quantify the uncertainty in an inferred field given the measurement of a related field when the two are linked by a mathematical model. Despite its many applications, Bayesian inference faces challenges when inferring fields that have discrete representations of large dimension, and/or have prior distributions that are difficult to characterize mathematically. In this work we demonstrate how the approximate distribution learned by a generative adversarial network (GAN) may be used as a prior in a Bayesian update

to address both these challenges. We demonstrate the efficacy of this approach by inferring and quantifying uncertainty in inference problems arising in computer vision and physics-based applications. In both instances we highlight the role of computing uncertainty in providing a measure of confidence in the solution, and in designing successive measurements to improve this confidence.
****

End to End Trainable Active Contours via Differentiable Rendering
Shir Gur,Tal Shaharabany,Lior Wolf
We present an image segmentation method that iteratively evolves a polygon. At each iteration, the vertices of the polygon are displaced based on the local value of a 2D shift map that is inferred from the input image via an encoder-decoder architecture. The main training loss that is used is the difference between the polygon shape and the ground truth segmentation mask. The network employs a neural renderer to create the polygon from its vertices, making the process fully differentiable. We demonstrate that our method outperforms the state of the art segmentation networks and deep active contour solutions in a variety of benchmarks, including medical imaging and aerial images.
****

Plan2Vec: Unsupervised Representation Learning by Latent Plans
Ge Yang,Amy Zhang,Ari Morcos,Joelle Pineau,Pieter Abbeel,Roberto Calandra
Creating a useful representation of the world takes more than just rote memorization of individual data samples. This is because fundamentally, we use our internal representation to plan, to solve problems, and to navigate the world. For a representation to be amenable to planning, it is critical for it to embody some notion of optimality. A representation learning objective that explicitly considers some form of planning should generate representations which are more computationally valuable than those that memorize samples. In this paper, we introduce \textbf{Plan2Vec}, an unsupervised representation learning objective inspired by value-based reinforcement learning methods. By abstracting away low-level control with a learned local metric, we show that it is possible to learn plannable representations that inform long-range structures, entirely passively from high-dimensional sequential datasets without supervision. A latent space is learned by playing an ``Imagined Planning Game" on the graph formed by the data points, using a local metric function trained contrastively from context. We show that the global metric on this learned embedding can be used to plan with O(1) complexity by linear interpolation. This exponential speed-up is critical for planning with a learned representation on any problem containing non-trivial global topology. We demonstrate the effectiveness of Plan2Vec on simulated toy tasks from both proprioceptive and image states, as well as two real-world image datasets, showing that Plan2Vec can effectively plan using learned representations. Additional results and videos can be found at \url{https://sites.google.com/view/plan2vec}.
****

Compositional Language Continual Learning
Yuanpeng Li,Liang Zhao,Kenneth Church,Mohamed Elhoseiny
Motivated by the human's ability to continually learn and gain knowledge over time, several research efforts have been pushing the limits of machines to constantly learn while alleviating catastrophic forgetting. Most of the existing methods have been focusing on continual learning of label prediction tasks, which have fixed input and output sizes. In this paper, we propose a new scenario of continual learning which handles sequence-to-sequence tasks common in language learning. We further propose an approach to use label prediction continual learning algorithm for sequence-to-sequence continual learning by leveraging compositionality. Experimental results show that the proposed method has significant improvement over state-of-the-art methods. It enables knowledge transfer and prevents catastrophic forgetting, resulting in more than 85% accuracy up to 100 stages, compared with less than 50% accuracy for baselines in instruction learning task. It also shows significant improvement in machine translation task. This is the first work to combine continual learning and compositionality for language learning, and we hope this work will make machines more helpful in various tasks.

****
Out-of-Distribution Image Detection Using the Normalized Compression Distance
Sehun Yu,Donga Lee,Hwanjo Yu
On detection of the out-of-distribution images, whose underlying distribution is different from that of the training dataset, we tackle to apply out-of-distribution detection methods to already deployed convolutional neural networks. Most recent approaches have to utilize out-of-distribution samples for validation or retrain the model, which makes it less practical for real-world applications. We propose a novel out-of-distribution detection method MALCOM, which neither uses any out-of-distribution samples nor retrain the model. Inspired by the method using the global average pooling on the feature maps of the convolutional neural networks, the goal of our method is to extract informative sequential patterns from the feature maps. To this end, we introduce a similarity metric which focuses on the shared patterns between two sequences. In short, MALCOM uses both the global average and spatial pattern of the feature maps to accurately identify out-of-distribution samples.
****
Discriminative Variational Autoencoder for Continual Learning with Generative Replay
Woo-Young Kang,Cheol-Ho Han,Byoung-Tak Zhang
Generative replay (GR) is a method to alleviate catastrophic forgetting in continual learning (CL) by generating previous task data and learning them together with the data from new tasks. In this paper, we propose discriminative variational autoencoder (DiVA) to address the GR-based CL problem. DiVA has class-wise discriminative latent embeddings by maximizing the mutual information between classes and latent variables of VAE. Thus, DiVA is directly applicable to classification and class-conditional generation which are efficient and effective properties in the GR-based CL scenario. Furthermore, we use a novel trick based on domain translation to cover natural images which is challenging to GR-based methods. As a result, DiVA achieved the competitive or higher accuracy compared to state-of-the-art algorithms in Permuted MNIST, Split MNIST, and Split CIFAR10 settings.
****
Connectivity-constrained interactive annotations for panoptic segmentation
Ruobing Shen,Bo Tang,Ismail Ben Ayed,Andrea Lodi,Thomas Guthier
Large-scale ground truth data sets are of crucial importance for deep learning based segmentation models, but annotating per-pixel
masks is prohibitively time consuming. In this paper, we investigate interactive graph-based segmentation algorithms that enforce connectivity. To be more precise, we introduce an instance-aware heuristic of a discrete Potts model, and a class-aware Integer Linear Programming (ILP) formulation that ensures global optimum. Both algorithms can take RGB, or utilize the feature maps from any DCNN, whether trained on the target dataset or not, as input. We present competitive semantic (and panoptic) segmentation results on the PASCAL VOC 2012 and Cityscapes dataset given initial scribbles. We also demonstrate that our interactive approach can reach 90.6% mIoU on VOC validation set with an overhead of just 3 correction scribbles. They are thus suitable for interactive annotation on new or existing datasets, or can be used inside any weakly supervised learning framework on new datasets.
****
Regularization Matters in Policy Optimization
Zhuang Liu,Xuanlin Li,Bingyi Kang,Trevor Darrell
Deep Reinforcement Learning (Deep RL) has been receiving increasingly more attention thanks to its encouraging performance on a variety of control tasks. Yet, conventional regularization techniques in training neural networks (e.g., L_2 regularization, dropout) have been largely ignored in RL methods, possibly because agents are typically trained and evaluated in the same environment. In this work, we present the first comprehensive study of regularization techniques with multiple policy optimization algorithms on continuous control tasks. Interestingly, we find conventional regularization techniques on the policy networks can often bring large improvement on the task performance, and the improvement is typi

cally more significant when the task is more difficult. We also compare with the widely used entropy regularization and find L_2 regularization is generally better. Our findings are further confirmed to be robust against the choice of training hyperparameters. We also study the effects of regularizing different components and find that only regularizing the policy network is typically enough. We hope our study provides guidance for future practices in regularizing policy optimization algorithms.

****

Adaptive Online Planning for Continual Lifelong Learning

Kevin Lu,Igor Mordatch,Pieter Abbeel

We study learning control in an online lifelong learning scenario, where mistakes can compound catastrophically into the future and the underlying dynamics of the environment may change. Traditional model-free policy learning methods have achieved successes in difficult tasks due to their broad flexibility, and capably condense broad experiences into compact networks, but struggle in this setting, as they can activate failure modes early in their lifetimes which are difficult to recover from and face performance degradation as dynamics change. On the other hand, model-based planning methods learn and adapt quickly, but require prohibitive levels of computational resources. Under constrained computation limits, the agent must allocate its resources wisely, which requires the agent to understand both its own performance and the current state of the environment: knowing that its mastery over control in the current dynamics is poor, the agent should dedicate more time to planning. We present a new algorithm, Adaptive Online Planning (AOP), that achieves strong performance in this setting by combining model-based planning with model-free learning. By measuring the performance of the planner and the uncertainty of the model-free components, AOP is able to call upon more extensive planning only when necessary, leading to reduced computations. We show that AOP gracefully deals with novel situations, adapting behaviors and policies effectively in the face of unpredictable changes in the world -- challenges that a continual learning agent naturally faces over an extended lifetime -- even when traditional reinforcement learning methods fail.

****

Measuring causal influence with back-to-back regression: the linear case

Jean-Remi King,Francois Charton,Maxime Oquab,David Lopez-Paz

Identifying causes from observations can be particularly challenging when i) potential factors are difficult to manipulate individually and ii) observations are complex and multi-dimensional. To address this issue, we introduce “Back-to-Back” regression (B2B), a method designed to efficiently measure, from a set of co-varying factors, the causal influences that most plausibly account for multidimensional observations. After proving the consistency of B2B and its links to other linear approaches, we show that our method outperforms least-squares regression and cross-decomposition techniques (e.g. canonical correlation analysis and partial least squares) on causal identification. Finally, we apply B2B to neuroimaging recordings of 102 subjects reading word sequences. The results show that the early and late brain representations, caused by low- and high-level word features respectively, are more reliably detected with B2B than with other standard techniques.

****

Multi-source Multi-view Transfer Learning in Neural Topic Modeling with Pretrained Topic and Word Embeddings

Pankaj Gupta,Yatin Chaudhary,Hinrich Schütze

Though word embeddings and topics are complementary representations, several past works have only used pretrained word embeddings in (neural) topic modeling to address data sparsity problem in short text or small collection of documents. However, no prior work has employed (pretrained latent) topics in transfer learning paradigm. In this paper, we propose a framework to perform transfer learning in neural topic modeling using (1) pretrained (latent) topics obtained from a large

source corpus, and (2) pretrained word and topic embeddings jointly (i.e., multi view)
in order to improve topic quality, better deal with polysemy and data sparsity issues in a target corpus. In doing so, we first accumulate topics and word repr esentations
from one or many source corpora to build respective pools of pretrained topic (i.e., TopicPool) and word embeddings (i.e., WordPool). Then, we identify one or multiple relevant source domain(s) and take advantage of corresponding topics and word features via the respective pools to guide meaningful learning in the sparse target domain. We quantify the quality of topic and document repre sentations
via generalization (perplexity), interpretability (topic coherence) and information retrieval (IR) using short-text, long-text, small and large document collections from news and medical domains. We have demonstrated the state-ofthe-art results on topic modeling with the proposed transfer learning approaches.
****

Adversarial Lipschitz Regularization
Dávid Terjék
Generative adversarial networks (GANs) are one of the most popular approaches wh en it comes to training generative models, among which variants of Wasserstein G ANs are considered superior to the standard GAN formulation in terms of learning stability and sample quality. However, Wasserstein GANs require the critic to b e 1-Lipschitz, which is often enforced implicitly by penalizing the norm of its gradient, or by globally restricting its Lipschitz constant via weight normaliza tion techniques. Training with a regularization term penalizing the violation of the Lipschitz constraint explicitly, instead of through the norm of the gradien t, was found to be practically infeasible in most situations. Inspired by Virtua l Adversarial Training, we propose a method called Adversarial Lipschitz Regular ization, and show that using an explicit Lipschitz penalty is indeed viable and leads to competitive performance when applied to Wasserstein GANs, highlighting an important connection between Lipschitz regularization and adversarial trainin g.
****

SGD Learns One-Layer Networks in WGANs
Qi Lei,Jason D. Lee,Alexandros G. Dimakis,Constantinos Daskalakis
Generative adversarial networks (GANs) are a widely used framework for learning generative models. Wasserstein GANs (WGANs), one of the most successful variants of GANs, require solving a minmax problem to global optimality, but in practice , are successfully trained with stochastic gradient descent-ascent. In this pape r, we show that, when the generator is a one-layer network, stochastic gradient descent-ascent converges to a global solution in polynomial time and sample comp lexity.
****

Localized Meta-Learning: A PAC-Bayes Analysis for Meta-Leanring Beyond Global Pr ior
Chenghao Liu,Tao Lu,Doyen Sahoo,Yuan Fang,Steven C.H. Hoi.
Meta-learning methods learn the meta-knowledge among various training tasks and aim to promote the learning of new tasks under the task similarity assumption. H owever, such meta-knowledge is often represented as a fixed distribution, which is too restrictive to capture various specific task information. In this work, w e present a localized meta-learning framework based on PAC-Bayes theory. In part icular, we propose a LCC-based prior predictor that allows the meta learner adap tively generate local meta-knowledge for specific task. We further develop a pra tical algorithm with deep neural network based on the bound. Empirical results o n real-world datasets demonstrate the efficacy of the proposed method.
****

Adversarial Training and Provable Defenses: Bridging the Gap
Mislav Balunovic,Martin Vechev
We present COLT, a new method to train neural networks based on a novel combinat ion of adversarial training and provable defenses. The key idea is to model neur

al network training as a procedure which includes both, the verifier and the adv ersary. In every iteration, the verifier aims to certify the network using conve x relaxation while the adversary tries to find inputs inside that convex relaxat ion which cause verification to fail. We experimentally show that this training method, named convex layerwise adversarial training (COLT), is promising and ach ieves the best of both worlds -- it produces a state-of-the-art neural network w ith certified robustness of 60.5% and accuracy of 78.4% on the challenging CIFAR -10 dataset with a 2/255 L-infinity perturbation. This significantly improves ov er the best concurrent results of 54.0% certified robustness and 71.5% accuracy.

****

Finding Deep Local Optima Using Network Pruning
Yangzi Guo,Yiyuan She,Ying Nian Wu,Adrian Barbu
Artificial neural networks (ANNs) are very popular nowadays and offer reliable s olutions to many classification problems. However, training deep neural networks (DNN) is time-consuming due to the large number of parameters. Recent research indicates that these DNNs might be over-parameterized and different solutions h ave been proposed to reduce the complexity both in the number of parameters and in the training time of the neural networks. Furthermore, some researchers argue that after reducing the neural network complexity via connection pruning, the r emaining weights are irrelevant and retraining the sub-network would obtain a co mparable accuracy with the original one.
This may hold true in most vision problems where we always enjoy a large number of training samples and research indicates that most local optima of the convolu tional neural networks may be equivalent. However, in non-vision sparse datasets , especially with many irrelevant features where a standard neural network would overfit, this might not be the case and there might be many non-equivalent loca l optima. This paper presents empirical evidence for these statements and an emp irical study of the learnability of neural networks (NNs) on some challenging no n-linear real and simulated data with irrelevant variables.
Our simulation experiments indicate that the cross-entropy loss function on XOR-like data has many local optima, and the number of local optima grows exponentia lly with the number of irrelevant variables.
We also introduce a connection pruning method to improve the capability of NNs t o find a deep local minimum even when there are irrelevant variables.
Furthermore, the performance of the discovered sparse sub-network degrades consi derably either by retraining from scratch or the corresponding original initiali zation, due to the existence of many bad optima around.
Finally, we will show that the performance of neural networks for real-world exp eriments on sparse datasets can be recovered or even improved by discovering a g ood sub-network architecture via connection pruning.
****

Adversarial Training Generalizes Data-dependent Spectral Norm Regularization
Kevin Roth,Yannic Kilcher,Thomas Hofmann
We establish a theoretical link between adversarial training and operator norm r egularization for deep neural networks. Specifically, we present a data-dependen t variant of spectral norm regularization and prove that it is equivalent to adv ersarial training based on a specific ℓ_2-norm constrained projected gradie nt ascent attack. This fundamental connection confirms the long-standing argumen t that a network's sensitivity to adversarial examples is tied to its spectral p roperties and hints at novel ways to robustify and defend against adversarial at tacks. We provide extensive empirical evidence to support our theoretical result s.
****

Knowledge Transfer via Student-Teacher Collaboration
Tianxiao Gao,Ruiqin Xiong,Zhenhua Liu,Siwei ma,Feng Wu,Tiejun Huang,Wen Gao
Accompanying with the flourish development in various fields, deep neural networ ks, however, are still facing with the plight of high computational costs and st orage. One way to compress these heavy models is knowledge transfer (KT), in whi

ch a light student network is trained through absorbing the knowledge from a pow erful teacher network. In this paper, we propose a novel knowledge transfer meth od which employs a Student-Teacher Collaboration (STC) network during the knowle dge transfer process. This is done by connecting the front part of the student n etwork to the back part of the teacher network as the STC network. The back part of the teacher network takes the intermediate representation from the front par t of the student network as input to make the prediction. The difference between the prediction from the collaboration network and the output tensor from the te acher network is taken into account of the loss during the train process. Throug h back propagation, the teacher network provides guidance to the student network in a gradient signal manner. In this way, our method takes advantage of the kno wledge from the entire teacher network, who instructs the student network in lea rning process. Through plentiful experiments, it is proved that our STC method o utperforms other KT methods with conventional strategy.
****

A Function Space View of Bounded Norm Infinite Width ReLU Nets: The Multivariate Case

Greg Ongie,Rebecca Willett,Daniel Soudry,Nathan Srebro

We give a tight characterization of the (vectorized Euclidean) norm of weights r equired to realize a function $f:\mathbb{R}\rightarrow \mathbb{R}^d$ as a single hidden-layer ReLU network with an unbounded number of units (infinite width), e xtending the univariate characterization of Savarese et al. (2019) to the multiv ariate case.
****

Weight-space symmetry in neural network loss landscapes revisited

Berfin Simsek,Johanni Brea,Bernd Illing,Wulfram Gerstner

Neural network training depends on the structure of the underlying loss landscap e, i.e. local minima, saddle points, flat plateaus, and loss barriers. In relati on to the structure of the landscape, we study the permutation symmetry of neuro ns in each layer of a deep neural network, which gives rise not only to multiple equivalent global minima of the loss function but also to critical points in be tween partner minima. In a network of $d-1$ hidden layers with n_k neurons in layers $k = 1, \ldots, d$, we construct continuous paths between equivalent glob al minima that lead through a `permutation point' where the input and output wei ght vectors of two neurons in the same hidden layer k collide and interchange. We show that such permutation points are critical points which lie inside high-dimensional subspaces of equal loss, contributing to the global flatness of the landscape. We also find that a permutation point for the exchange of neurons i and j transits into a flat high-dimensional plateau that enables all $n_k!$ p ermutations of neurons in a given layer k at the same loss value. Moreover, we introduce higher-order permutation points by exploiting the hierarchical struct ure in the loss landscapes of neural networks, and find that the number of K-t h order permutation points is much larger than the (already huge) number of equi valent global minima -- at least by a polynomial factor of order K. In two tas ks, we demonstrate numerically with our path finding method that continuous path s between partner minima exist: first, in a toy network with a single hidden lay er on a function approximation task and, second, in a multilayer network on the MNIST task. Our geometric approach yields a lower bound on the number of critica l points generated by weight-space symmetries and provides a simple intuitive li nk between previous theoretical results and numerical observations.
****

Differentiable Bayesian Neural Network Inference for Data Streams

Namuk Park,Taekyu Lee,Songkuk Kim

While deep neural networks (NNs) do not provide the confidence of its prediction , Bayesian neural network (BNN) can estimate the uncertainty of the prediction. However, BNNs have not been widely used in practice due to the computational co st of predictive inference. This prohibitive computational cost is a hindrance e specially when processing stream data with low-latency. To address this problem, we propose a novel model which approximate BNNs for data streams. Instead of ge nerating separate prediction for each data sample independently, this model esti

mates the increments of prediction for a new data sample from the previous predi
ctions. The computational cost of this model is almost the same as that of non-B
ayesian deep NNs. Experiments including semantic segmentation on real-world data
 show that this model performs significantly faster than BNNs, estimating uncert
ainty comparable to the results of BNNs.

****
Lite Transformer with Long-Short Range Attention
Zhanghao Wu*,Zhijian Liu*,Ji Lin,Yujun Lin,Song Han
Transformer has become ubiquitous in natural language processing (e.g., machine
 translation, question answering); however, it requires enormous amount of comput
ations to achieve high performance, which makes it not suitable for mobile appli
cations that are tightly constrained by the hardware resources and battery. In t
his paper, we present an efficient mobile NLP architecture, Lite Transformer to
facilitate deploying mobile NLP applications on edge devices. The key primitive
is the Long-Short Range Attention (LSRA), where one group of heads specializes i
n the local context modeling (by convolution) while another group specializes in
 the long-distance relationship modeling (by attention). Such specialization bri
ngs consistent improvement over the vanilla transformer on three well-establishe
d language tasks: machine translation, abstractive summarization, and language m
odeling. Under constrained resources (500M/100M MACs), Lite Transformer outperfo
rms transformer on WMT'14 English-French by 1.2/1.7 BLEU, respectively. Lite Tra
nsformer reduces the computation of transformer base model by 2.5x with 0.3 BLEU
 score degradation. Combining with pruning and quantization, we further compress
ed the model size of Lite Transformer by 18.2x. For language modeling, Lite Tran
sformer achieves 1.8 lower perplexity than the transformer at around 500M MACs.
Notably, Lite Transformer outperforms the AutoML-based Evolved Transformer by 0.
5 higher BLEU for the mobile NLP setting without the costly architecture search
that requires more than 250 GPU years. Code has been made available at https://g
ithub.com/mit-han-lab/lite-transformer.
****
Learning by shaking: Computing policy gradients by physical forward-propagation
Arash Mehrjou,Ashkan Soleymani,Stefan Bauer,Bernhard Schölkopf
Model-free and model-based reinforcement learning are two ends of a spectrum. Le
arning a good policy without a dynamic model can be prohibitively expensive. Lea
rning the dynamic model of a system can reduce the cost of learning the policy,
but it can also introduce bias if it is not accurate. We propose a middle ground
 where instead of the transition model, the sensitivity of the trajectories with
 respect to the perturbation (shaking) of the parameters is learned. This allows
 us to predict the local behavior of the physical system around a set of nominal
 policies without knowing the actual model. We assay our method on a custom-buil
t physical robot in extensive experiments and show the feasibility of the approa
ch in practice. We investigate potential challenges when applying our method to
physical systems and propose solutions to each of them.
****
Occlusion resistant learning of intuitive physics from videos
Ronan Riochet,Josef Sivic,Ivan Laptev,Emmanuel Dupoux
To reach human performance on complex tasks, a key ability for artificial
systems is to understand physical interactions between objects, and predict
future outcomes of a situation. This ability, often referred to as
intuitive
physics
, has recently received attention and several methods were proposed to
learn these physical rules from video sequences. Yet, most these methods are
restricted to the case where no occlusions occur, narrowing the potential areas
of application. The main contribution of this paper is a method combining
a predictor of object dynamics and a neural renderer efficiently predicting
future trajectories and explicitly modelling partial and full occlusions among
objects. We present a training procedure enabling learning intuitive physics
directly from the input videos containing segmentation masks of objects and

their depth. Our results show that our model learns object dynamics despite significant inter-object occlusions, and realistically predicts segmentation masks up to 30 frames in the future. We study model performance for increasing levels of occlusions, and compare results to previous work on the tasks of future prediction and object following. We also show results on predicting motion of objects in real videos and demonstrate significant improvements over state-of-the-art on the object permanence task in the intuitive physics benchmark of Riochet et al. (2018).
****
Statistical Verification of General Perturbations by Gaussian Smoothing
Marc Fischer,Maximilian Baader,Martin Vechev
We present a novel statistical certification method that generalizes prior work based on smoothing to handle richer perturbations. Concretely, our method produces a provable classifier which can establish statistical robustness against geometric perturbations (e.g., rotations, translations) as well as volume changes and pitch shifts on audio data. The generalization is non-trivial and requires careful handling of operations such as interpolation. Our method is agnostic to the choice of classifier and scales to modern architectures such as ResNet-50 on ImageNet.
****
Localised Generative Flows
Rob Cornish,Anthony Caterini,George Deligiannidis,Arnaud Doucet
We argue that flow-based density models based on continuous bijections are limited in their ability to learn target distributions with complicated topologies, and propose localised generative flows (LGFs) to address this problem. LGFs are composed of stacked continuous mixtures of bijections, which enables each bijection to learn a local region of the target rather than its entirety. Our method is a generalisation of existing flow-based methods, which can be used without modification as the basis for an LGF model. Unlike normalising flows, LGFs do not permit exact computation of log likelihoods, but we propose a simple variational scheme that performs well in practice. We show empirically that LGFs yield improved performance across a variety of common density estimation tasks.
****
Certifying Neural Network Audio Classifiers
Wonryong Ryou,Mislav Balunovic,Gagandeep Singh,Martin Vechev
We present the first end-to-end verifier of audio classifiers. Compared to existing methods, our approach enables analysis of both, the entire audio processing stage as well as recurrent neural network architectures (e.g., LSTM). The audio processing is verified using novel convex relaxations tailored to feature extraction operations used in audio (e.g., Fast Fourier Transform) while recurrent architectures are certified via a novel binary relaxation for the recurrent unit update. We show the verifier scales to large networks while computing significantly tighter bounds than existing methods for common audio classification benchmarks: on the challenging Google Speech Commands dataset we certify 95% more inputs than the interval approximation (only prior scalable method), for a perturbation of -90dB.
****
Collaborative Training of Balanced Random Forests for Open Set Domain Adaptation
Jongbin Ryu,Jiun Bae,Jongwoo Lim
In this paper, we introduce a collaborative training algorithm of balanced random forests for domain adaptation tasks which can avoid the overfitting problem. In real scenarios, most domain adaptation algorithms face the challenges from noisy, insufficient training data. Moreover in open set categorization, unknown or misaligned source and target categories adds difficulty. In such cases, conventional methods suffer from overfitting and fail to successfully transfer the knowledge of the source to the target domain. To address these issues, the following two techniques are proposed. First, we introduce the optimized decision tree construction method, in which the data at each node are split into equal sizes while maximizing the information gain. Compared to the conventional random forests, it generates larger and more balanced decision trees due to the even-split const

raint, which contributes to enhanced discrimination power and reduced overfittin g. Second, to tackle the domain misalignment problem, we propose the domain alig nment loss which penalizes uneven splits of the source and target domain data. B y collaboratively optimizing the information gain of the labeled source data as well as the entropy of unlabeled target data distributions, the proposed CoBRF a lgorithm achieves significantly better performance than the state-of-the-art met hods. The proposed algorithm is extensively evaluated in various experimental se tups in challenging domain adaptation tasks with noisy and small training data a s well as open set domain adaptation problems, for two backbone networks of Alex Net and ResNet-50.
****

PAC-Bayesian Neural Network Bounds
Yossi Adi,Alex Schwing,Tamir Hazan
Bayesian neural networks, which both use the negative log-likelihood loss functi on and average their predictions using a learned posterior over the parameters, have been used successfully across many scientific fields, partly due to their a bility to `effortlessly' extract desired representations from many large-scale d atasets. However, generalization bounds for this setting is still missing.
In this paper, we present a new PAC-Bayesian generalization bound for the negati ve log-likelihood loss which utilizes the \emph{Herbst Argument} for the log-Sob olev inequality to bound the moment generating function of the learners risk.
****

Semi-Implicit Back Propagation
Ren Liu,Xiaoqun Zhang
Neural network has attracted great attention for a long time and many researcher s are devoted to improve the effectiveness of neural network training algorith ms. Though stochastic gradient descent (SGD) and other explicit gradient-based m ethods are widely adopted, there are still many challenges such as gradient vani shing and small step sizes, which leads to slow convergence and instability of S GD algorithms. Motivated by error back propagation (BP) and proximal methods, w e propose a semi-implicit back propagation method for neural network training. S imilar to BP, the difference on the neurons are propagated in a backward fashion and the parameters are updated with proximal mapping. The implicit update for b oth hidden neurons and parameters allows to choose large step size in the traini ng algorithm. Finally, we also show that any fixed point of convergent sequence s produced by this algorithm is a stationary point of the objective loss functio n. The experiments on both MNIST and CIFAR-10 demonstrate that the proposed semi -implicit BP algorithm leads to better performance in terms of both loss decrea sing and training/validation accuracy, compared to SGD and a similar algorithm P roxBP.
****

Mutual Information Gradient Estimation for Representation Learning
Liangjian Wen,Yiji Zhou,Lirong He,Mingyuan Zhou,Zenglin Xu
Mutual Information (MI) plays an important role in representation learning. Howe ver, MI is unfortunately intractable in continuous and high-dimensional settings . Recent advances establish tractable and scalable MI estimators to discover use ful representation. However, most of the existing methods are not capable of pro viding an accurate estimation of MI with low-variance when the MI is large. We a rgue that directly estimating the gradients of MI is more appealing for represen tation learning than estimating MI in itself. To this end, we propose the Mutual Information Gradient Estimator (MIGE) for representation learning based on the score estimation of implicit distributions. MIGE exhibits a tight and smooth gra dient estimation of MI in the high-dimensional and large-MI settings. We expand the applications of MIGE in both unsupervised learning of deep representations b ased on InfoMax and the Information Bottleneck method. Experimental results have indicated significant performance improvement in learning useful representation .
****

Qgraph-bounded Q-learning: Stabilizing Model-Free Off-Policy Deep Reinforcement Learning

Sabrina Hoppe,Marc Toussaint
In state of the art model-free off-policy deep reinforcement learning (RL), a replay memory is used to store past experience and derive all network updates. Even if both state and action spaces are continuous, the replay memory only holds a finite number of transitions. We represent these transitions in a data graph and link its structure to soft divergence. By selecting a subgraph with a favorable structure, we construct a simple Markov Decision Process (MDP) for which exact Q-values can be computed efficiently as more data comes in - resulting in a Qgraph. We show that the Q-value for each transition in the simplified MDP is a lower bound of the Q-value for the same transition in the original continuous Q-learning problem. By using these lower bounds in TD learning, our method is less prone to soft divergence and exhibits increased sample efficiency while being more robust to hyperparameters. Qgraphs also retain information from transitions that have already been overwritten in the replay memory, which can decrease the algorithm's sensitivity to the replay memory capacity.

****
Iterative Deep Graph Learning for Graph Neural Networks
Yu Chen,Lingfei Wu,Mohammed J. Zaki
In this paper, we propose an end-to-end graph learning framework, namely Iterative Deep Graph Learning (IDGL), for jointly learning graph structure and graph embedding simultaneously. We first cast graph structure learning problem as similarity metric learning problem and leverage an adapted graph regularization for controlling smoothness, connectivity and sparsity of the generated graph. We further propose a novel iterative method for searching for hidden graph structure that augments the initial graph structure. Our iterative method dynamically stops when learning graph structure approaches close enough to the ground truth graph. Our extensive experiments demonstrate that the proposed IDGL model can consistently outperform or match state-of-the-art baselines in terms of both classification accuracy and computational time. The proposed approach can cope with both transductive training and inductive training.
****
Mint: Matrix-Interleaving for Multi-Task Learning
Tianhe Yu,Saurabh Kumar,Eric Mitchell,Abhishek Gupta,Karol Hausman,Sergey Levine,Chelsea Finn
Deep learning enables training of large and flexible function approximators from scratch at the cost of large amounts of data. Applications of neural networks often consider learning in the context of a single task. However, in many scenarios what we hope to learn is not just a single task, but a model that can be used to solve multiple different tasks. Such multi-task learning settings have the potential to improve data efficiency and generalization by sharing data and representations across tasks. However, in some challenging multi-task learning settings, particularly in reinforcement learning, it is very difficult to learn a single model that can solve all the tasks while realizing data efficiency and performance benefits. Learning each of the tasks independently from scratch can actually perform better in such settings, but it does not benefit from the representation sharing that multi-task learning can potentially provide. In this work, we develop an approach that endows a single model with the ability to represent both extremes: joint training and independent training. To this end, we introduce matrix-interleaving (Mint), a modification to standard neural network models that projects the activations for each task into a different learned subspace, represented by a per-task and per-layer matrix. By learning these matrices jointly with the other model parameters, the optimizer itself can decide how much to share representations between tasks. On three challenging multi-task supervised learning and reinforcement learning problems with varying degrees of shared task structure, we find that this model consistently matches or outperforms joint training and independent training, combining the best elements of both.
****
Learning Cluster Structured Sparsity by Reweighting
Yulun Jiang,Lei Yu,Haijian Zhang,Zhou Liu

Recently, the paradigm of unfolding iterative algorithms into finite-length feed -forward neural networks has achieved a great success in the area of sparse reco very. Benefit from available training data, the learned networks have achieved s tate-of-the-art performance in respect of both speed and accuracy. However, the structure behind sparsity, imposing constraint on the support of sparse signals, is often an essential prior knowledge but seldom considered in the existing net works. In this paper, we aim at bridging this gap. Specifically, exploiting the iterative reweighted ℓ_1 minimization (IRL1) algorithm, we propose to learn the cluster structured sparsity (CSS) by rewegihting adaptively. In particular, we first unfold the Reweighted Iterative Shrinkage Algorithm (RwISTA) into an e nd-to-end trainable deep architecture termed as RW-LISTA. Then instead of the el ement-wise reweighting, the global and local reweighting manner are proposed for the cluster structured sparse learning. Numerical experiments further show the superiority of our algorithm against both classical algorithms and learning-base d networks on different tasks.

****

Selfish Emergent Communication

Michael Noukhovitch,Travis LaCroix,Aaron Courville

Current literature in machine learning holds that unaligned, self-interested age nts do not learn to use an emergent communication channel. We introduce a new se nder-receiver game to study emergent communication for this spectrum of partiall y-competitive scenarios and put special care into evaluation. We find that commu nication can indeed emerge in partially-competitive scenarios, and we discover t hree things that are tied to improving it. First, that selfish communication is proportional to cooperation, and it naturally occurs for situations that are mor e cooperative than competitive. Second, that stability and performance are impro ved by using LOLA (Foerster et al, 2018), especially in more competitive scenari os. And third, that discrete protocols lend themselves better to learning cooper ative communication than continuous ones.

****

Decoupling Adaptation from Modeling with Meta-Optimizers for Meta Learning

Sébastien M.R. Arnold,Shariq Iqbal,Fei Sha

Meta-learning methods, most notably Model-Agnostic Meta-Learning (Finn et al, 20 17) or MAML, have achieved great success in adapting to new tasks quickly, after having been trained on similar tasks.

The mechanism behind their success, however, is poorly understood.

We begin this work with an experimental analysis of MAML, finding that deep mode ls are crucial for its success, even given sets of simple tasks where a linear m odel would suffice on any individual task.

Furthermore, on image-recognition tasks, we find that the early layers of MAML-t rained models learn task-invariant features, while later layers are used for ada ptation, providing further evidence that these models require greater capacity t han is strictly necessary for their individual tasks.

Following our findings, we propose a method which enables better use of model ca pacity at inference time by separating the adaptation aspect of meta-learning in to parameters that are only used for adaptation but are not part of the forward model.

We find that our approach enables more effective meta-learning in smaller models , which are suitably sized for the individual tasks.

****

Imitation Learning of Robot Policies using Language, Vision and Motion

Simon Stepputtis,Joseph Campbell,Mariano Phielipp,Chitta Baral,Heni Ben Amor

In this work we propose a novel end-to-end imitation learning approach which com bines natural language, vision, and motion information to produce an abstract re presentation of a task, which in turn can be used to synthesize specific motion controllers at run-time. This multimodal approach enables generalization to a wi de variety of environmental conditions and allows an end-user to influence a rob ot policy through verbal communication. We empirically validate our approach wit h an extensive set of simulations and show that it achieves a high task success

rate over a variety of conditions while remaining amenable to probabilistic inte
rpretability.
****

Improving Visual Relation Detection using Depth Maps
Sahand Sharifzadeh,Sina Moayed Baharlou,Max Berrendorf,Rajat Koner,Volker Tresp
State of the art visual relation detection methods mostly rely on object informa
tion extracted from RGB images such as predicted class probabilities, 2D boundin
g boxes and feature maps. In this paper, we argue that the 3D positions of objec
ts in space can provide additional valuable information about object relations.
This information helps not only to detect spatial relations, such as \textit{sta
nding behind}, but also non-spatial relations, such as \textit{holding}. Since 3
D information of a scene is not easily accessible, we propose incorporating a pr
e-trained RGB-to-Depth model within visual relation detection frameworks. We dis
cuss different feature extraction strategies from depth maps and show their crit
ical role in relation detection.
Our experiments confirm that the performance of state-of-the-art visual relation
 detection approaches can significantly be improved by utilizing depth map infor
mation.
****

Semi-supervised Pose Estimation with Geometric Latent Representations
Luis A. Perez Rey,Dmitri Jarnikov,Mike Holenderski
Pose estimation is the task of finding the orientation of an object within an i
mage with respect to a fixed frame of reference. Current classification and regr
ession approaches to the task require large quantities of labelled data for thei
r purposes. The amount of labelled data for pose estimation is relatively limite
d. With this in mind, we propose the use of Conditional Variational Autoencoders
 (CVAEs) \cite{Kingma2014a} with circular latent representations to estimate the
 corresponding 2D rotations of an object. The method is capable of training with
 datasets that have an arbitrary amount of labelled images providing relatively
similar performance for cases in which 10-20% of the labels for images is missin
g.
****

Identifying Weights and Architectures of Unknown ReLU Networks
David Rolnick,Konrad P. Kording
The output of a neural network depends on its parameters in a highly nonlinear w
ay, and it is widely assumed that a network's parameters cannot be identified fr
om its outputs. Here, we show that in many cases it is possible to reconstruct t
he architecture, weights, and biases of a deep ReLU network given the ability to
 query the network. ReLU networks are piecewise linear and the boundaries betwee
n pieces correspond to inputs for which one of the ReLUs switches between inacti
ve and active states. Thus, first-layer ReLUs can be identified (up to sign and
scaling) based on the orientation of their associated hyperplanes. Later-layer R
eLU boundaries bend when they cross earlier-layer boundaries and the extent of b
ending reveals the weights between them. Our algorithm uses this to identify the
 units in the network and weights connecting them (up to isomorphism). The fact
that considerable parts of deep networks can be identified from their outputs ha
s implications for security, neuroscience, and our understanding of neural netwo
rks.
****

Unsupervised Domain Adaptation through Self-Supervision
Yu Sun,Eric Tzeng,Trevor Darrell,Alexei A. Efros
This paper addresses unsupervised domain adaptation, the setting where labeled t
raining data is available on a source domain, but the goal is to have good perfo
rmance on a target domain with only unlabeled data. Like much of previous work,
we seek to align the learned representations of the source and target domains wh
ile preserving discriminability. The way we accomplish alignment is by learning
to perform auxiliary self-supervised task(s) on both domains simultaneously. Ea
ch self-supervised task brings the two domains closer together along the directi
on relevant to that task. Training this jointly with the main task classifier on
 the source domain is shown to successfully generalize to the unlabeled target d

omain. The presented objective is straightforward to implement and easy to opti
mize. We achieve state-of-the-art results on four out of seven standard benchmar
ks, and competitive results on segmentation adaptation. We also demonstrate that
 our method composes well with another popular pixel-level adaptation method.
****
Improving Gradient Estimation in Evolutionary Strategies With Past Descent Direc
tions
Florian Meier,Asier Mujika,Marcelo Gauy,Angelika Steger
We propose a novel method to optimally incorporate surrogate gradient informatio
n. Our approach, unlike previous work, needs no information about the quality o
f the surrogate gradients and is always guaranteed to find a descent direction t
hat is better than the surrogate gradient. This allows to iteratively use the p
revious gradient estimate as surrogate gradient for the current search point. We
 theoretically prove that this yields fast convergence to the true gradient for
 linear functions and show under simplifying assumptions that it significantly i
mproves gradient estimates for general functions. Finally, we evaluate our appro
ach empirically on MNIST and reinforcement learning tasks and show that it consi
derably improves the gradient estimation of ES at no extra computational cost.
****
Variable Complexity in the Univariate and Multivariate Structural Causal Model
Tomer Galanti,Ofir Nabati,Lior Wolf
We show that by comparing the individual complexities of univariante cause and e
ffect in the Structural Causal Model, one can identify the cause and the effect,
 without considering their interaction at all. The entropy of each variable is i
neffective in measuring the complexity, and we propose to capture it by an autoe
ncoder that operates on the list of sorted samples. Comparing the reconstruction
 errors of the two autoencoders, one for each variable, is shown to perform well
 on the accepted benchmarks of the field.

In the multivariate case, where one can ensure that the complexities of the caus
e and effect are balanced, we propose a new method that mimics the disentangled
structure of the causal model. We extend the results of~\cite{Zhang:2009:IPC:179
5114.1795190} to the multidimensional case, showing that such modeling is only l
ikely in the direction of causality. Furthermore, the learned model is shown the
oretically to perform the separation to the causal component and to the residual
  (noise) component. Our multidimensional method obtains a significantly higher
accuracy than the literature methods.
****
Regularizing activations in neural networks via distribution matching with the W
asserstein metric
Taejong Joo,Donggu Kang,Byunghoon Kim
Regularization and normalization have become indispensable components in trainin
g deep neural networks, resulting in faster training and improved generalization
 performance. We propose the projected error function regularization loss (PER)
that encourages activations to follow the standard normal distribution. PER rand
omly projects activations onto one-dimensional space and computes the regulariza
tion loss in the projected space. PER is similar to the Pseudo-Huber loss in the
 projected space, thus taking advantage of both L^1 and L^2 regularization l
osses. Besides, PER can capture the interaction between hidden units by projecti
on vector drawn from a unit sphere. By doing so, PER minimizes the upper bound o
f the Wasserstein distance of order one between an empirical distribution of act
ivations and the standard normal distribution. To the best of the authors' knowl
edge, this is the first work to regularize activations via distribution matching
 in the probability distribution space. We evaluate the proposed method on the i
mage classification task and the word-level language modeling task.

****
Gradient Descent Maximizes the Margin of Homogeneous Neural Networks
Kaifeng Lyu,Jian Li
In this paper, we study the implicit regularization of the gradient descent algo

rithm in homogeneous neural networks, including fully-connected and convolutiona l neural networks with ReLU or LeakyReLU activations. In particular, we study th e gradient descent or gradient flow (i.e., gradient descent with infinitesimal s tep size) optimizing the logistic loss or cross-entropy loss of any homogeneous model (possibly non-smooth), and show that if the training loss decreases below a certain threshold, then we can define a smoothed version of the normalized mar gin which increases over time. We also formulate a natural constrained optimizat ion problem related to margin maximization, and prove that both the normalized m argin and its smoothed version converge to the objective value at a KKT point of the optimization problem. Our results generalize the previous results for logis tic regression with one-layer or multi-layer linear networks, and provide more q uantitative convergence results with weaker assumptions than previous results fo r homogeneous smooth neural networks. We conduct several experiments to justify our theoretical finding on MNIST and CIFAR-10 datasets. Finally, as margin is cl osely related to robustness, we discuss potential benefits of training longer fo r improving the robustness of the model.
****

Mixed Precision Training With 8-bit Floating Point
Naveen Mellempudi,Sudarshan Srinivasan,Dipankar Das,Bharat Kaul
Reduced precision computation is one of the key areas addressing the widening’co mpute gap’, driven by an exponential growth in deep learning applications. In re cent years, deep neural network training has largely migrated to 16-bit precisio n,with significant gains in performance and energy efficiency. However, attempts to train DNNs at 8-bit precision have met with significant challenges, because of the higher precision and dynamic range requirements of back-propagation. In this paper, we propose a method to train deep neural networks using 8-bit floating point representation for weights, activations, errors, and gradi ents. We demonstrate state-of-the-art accuracy across multiple data sets (image net-1K, WMT16)and a broader set of workloads (Resnet-18/34/50, GNMT, and Transfo rmer) than previously reported. We propose an enhanced loss scaling method to augment the reduced subnormal range of 8-bit floating point, to improve error pr opagation.We also examine the impact of quantization noise on generalization, an d propose a stochastic rounding technique to address gradient noise. As a result of applying all these techniques, we report slightly higher validation accurac y compared to full precision baseline.
****

An Empirical and Comparative Analysis of Data Valuation with Scalable Algorithms
Ruoxi Jia,Xuehui Sun,Jiacen Xu,Ce Zhang,Bo Li,Dawn Song
This paper focuses on valuating training data for supervised learning tasks and studies the Shapley value, a data value notion originated in cooperative game th eory. The Shapley value defines a unique value distribution scheme that satisfie s a set of appealing properties desired by a data value notion. However, the Sha pley value requires exponential complexity to calculate exactly. Existing approx imation algorithms, although achieving great improvement over the exact algorith m, relies on retraining models for multiple times, thus remaining limited when a pplied to larger-scale learning tasks and real-world datasets.

In this work, we develop a simple and efficient algorithm to estimate the Shaple y value with complexity independent with the model size. The key idea is to appr oximate the model via a K-nearest neighbor (KNN) classifier, which has a loc ality structure that can lead to efficient Shapley value calculation. We evaluat e the utility of the values produced by the KNN proxies in various settings, i ncluding label noise correction, watermark detection, data summarization, active data acquisition, and domain adaption. Extensive experiments demonstrate that o ur algorithm achieves at least comparable utility to the values produced by exis ting algorithms while significant efficiency improvement. Moreover, we theoretic ally analyze the Shapley value and justify its advantage over the leave-one-out error as a data value measure.
****

Consistent Meta-Reinforcement Learning via Model Identification and Experience R

elabeling
Russell Mendonca,Xinyang Geng,Chelsea Finn,Sergey Levine
Reinforcement learning algorithms can acquire policies for complex tasks automat
ically, however the number of samples required to learn a diverse set of skills
can be prohibitively large. While meta-reinforcement learning has enabled agents
 to leverage prior experience to adapt quickly to new tasks, the performance of
these methods depends crucially on how close the new task is to the previously e
xperienced tasks. Current approaches are either not able to extrapolate well, or
 can do so at the expense of requiring extremely large amounts of data due to on
-policy training. In this work, we present model identification and experience r
elabeling (MIER), a meta-reinforcement learning algorithm that is both efficient
 and extrapolates well when faced with out-of-distribution tasks at test time ba
sed on a simple insight: we recognize that dynamics models can be adapted effici
ently and consistently with off-policy data, even if policies and value function
s cannot. These dynamics models can then be used to continue training policies f
or out-of-distribution tasks without using meta-reinforcement learning at all, b
y generating synthetic experience for the new task.
****

Transferring Optimality Across Data Distributions via Homotopy Methods
Matilde Gargiani,Andrea Zanelli,Quoc Tran Dinh,Moritz Diehl,Frank Hutter
Homotopy methods, also known as continuation methods, are a powerful mathematica
l tool to efficiently solve various problems in numerical analysis, including co
mplex non-convex optimization problems where no or only little prior knowledge r
egarding the localization of the solutions is available.
In this work, we propose a novel homotopy-based numerical method that can be use
d to transfer knowledge regarding the localization of an optimum across differen
t task distributions in deep learning applications. We validate the proposed met
hodology with some empirical evaluations in the regression and classification sc
enarios, where it shows that superior numerical performance can be achieved in p
opular deep learning benchmarks, i.e. FashionMNIST, CIFAR-10, and draw connectio
ns with the widely used fine-tuning heuristic. In addition, we give more insight
s on the properties of a general homotopy method when used in combination with S
tochastic Gradient Descent by conducting a general local theoretical analysis in
 a simplified setting.
****

Latent Normalizing Flows for Many-to-Many Cross-Domain Mappings
Shweta Mahajan,Iryna Gurevych,Stefan Roth
Learned joint representations of images and text form the backbone of several im
portant cross-domain tasks such as image captioning. Prior work mostly maps both
 domains into a common latent representation in a purely supervised fashion. Thi
s is rather restrictive, however, as the two domains follow distinct generative
processes. Therefore, we propose a novel semi-supervised framework, which models
 shared information between domains and domain-specific information separately.
The information shared between the domains is aligned with an invertible neural
network. Our model integrates normalizing flow-based priors for the domain-speci
fic information, which allows us to learn diverse many-to-many mappings between
the two domains. We demonstrate the effectiveness of our model on diverse tasks,
 including image captioning and text-to-image synthesis.
****

Dynamic Model Pruning with Feedback
Tao Lin,Sebastian U. Stich,Luis Barba,Daniil Dmitriev,Martin Jaggi
Deep neural networks often have millions of parameters. This can hinder their de
ployment to low-end devices, not only due to high memory requirements but also b
ecause of increased latency at inference. We propose a novel model compression m
ethod that generates a sparse trained model without additional overhead: by allo
wing (i) dynamic allocation of the sparsity pattern and (ii) incorporating feed
back signal to reactivate prematurely pruned weights we obtain a performant spar
se model in one single training pass (retraining is not needed, but can further
improve the performance). We evaluate the method on CIFAR-10 and ImageNet, and s
how that the obtained sparse models can reach the state-of-the-art performance o

f dense models and further that their performance surpasses all previously propo
sed pruning schemes (that come without feedback mechanisms).
****

ℓ_1 Adversarial Robustness Certificates: a Randomized Smoothing Approach
Jiaye Teng,Guang-He Lee,Yang Yuan
Robustness is an important property to guarantee the security of machine learnin
g models. It has recently been demonstrated that strong robustness certificates
can be obtained on ensemble classifiers generated by input randomization. Howeve
r, tight robustness certificates are only known for symmetric norms including $\
ell_0$ and ℓ_2, while for asymmetric norms like ℓ_1, the existing tech
niques do not apply. By converting the likelihood ratio into a one-dimensional m
ixed random variable, we derive the first tight ℓ_1 robustness certificate
under isotropic Laplace distributions. Empirically, the deep networks smoothed b
y Laplace distributions yield the state-of-the-art certified robustness in $\ell
_1$ norm on CIFAR-10 and ImageNet.
****

On the interaction between supervision and self-play in emergent communication
Ryan Lowe*,Abhinav Gupta*,Jakob Foerster,Douwe Kiela,Joelle Pineau
A promising approach for teaching artificial agents to use natural language invo
lves using human-in-the-loop training. However, recent work suggests that curren
t machine learning methods are too data inefficient to be trained in this way fr
om scratch. In this paper, we investigate the relationship between two categorie
s of learning signals with the ultimate goal of improving sample efficiency: imi
tating human language data via supervised learning, and maximizing reward in a s
imulated multi-agent environment via self-play (as done in emergent communicatio
n), and introduce the term supervised self-play (S2P) for algorithms using both
of these signals. We find that first training agents via supervised learning on
human data followed by self-play outperforms the converse, suggesting that it is
 not beneficial to emerge languages from scratch. We then empirically investigat
e various S2P schedules that begin with supervised learning in two environments:
 a Lewis signaling game with symbolic inputs, and an image-based referential gam
e with natural language descriptions. Lastly, we introduce population based appr
oaches to S2P, which further improves the performance over single-agent methods.
****

CNAS: Channel-Level Neural Architecture Search
Heechul Lim,Min-Soo Kim,Jinjun Xiong
There is growing interest in automating designing good neural network architectu
res. The NAS methods proposed recently have significantly reduced architecture s
earch cost by sharing parameters, but there is still a challenging problem of de
signing search space. We consider search space is typically defined with its sha
pe and a set of operations and propose a channel-level architecture search\,(CNA
S) method using only a fixed type of operation. The resulting architecture is sp
arse in terms of channel and has different topology at different cell. The exper
imental results for CIFAR-10 and ImageNet show that a fine-granular and sparse m
odel searched by CNAS achieves very competitive performance with dense models se
arched by the existing methods.
****

FLAT MANIFOLD VAES
Nutan Chen,Alexej Klushyn,Francesco Ferroni,Justin Bayer,Patrick van der Smagt
Latent-variable models represent observed data by mapping a prior distribution o
ver some latent space to an observed space. Often, the prior distribution is sp
ecified by the user to be very simple, effectively shifting the burden of a lear
ning algorithm to the estimation of a highly non-linear likelihood function. Thi
s poses a problem for the calculation of a popular distance function, the geodes
ic between data points in the latent space, as this is often solved iteratively
via numerical methods. These are less effective if the problem at hand is not we
ll captured by first or second-order approximations. In this work, we propose le
ss complex likelihood functions by allowing complex distributions and explicitly
 penalising the curvature of the decoder. This results in geodesics which are ap
proximated well by the Euclidean distance in latent space, decreasing the runtim

e by a factor of 1,000 with little loss in accuracy.

****
A Meta-Transfer Objective for Learning to Disentangle Causal Mechanisms
Yoshua Bengio,Tristan Deleu,Nasim Rahaman,Nan Rosemary Ke,Sebastien Lachapelle,Olexa Bilaniuk,Anirudh Goyal,Christopher Pal
We propose to use a meta-learning objective that maximizes the speed of transfer on a modified distribution to learn how to modularize acquired knowledge. In particular, we focus on how to factor a joint distribution into appropriate conditionals, consistent with the causal directions. We explain when this can work, using the assumption that the changes in distributions are localized (e.g. to one of the marginals, for example due to an intervention on one of the variables). We prove that under this assumption of localized changes in causal mechanisms, the correct causal graph will tend to have only a few of its parameters with non-zero gradient, i.e. that need to be adapted (those of the modified variables). We argue and observe experimentally that this leads to faster adaptation, and use this property to define a meta-learning surrogate score which, in addition to a continuous parametrization of graphs, would favour correct causal graphs. Finally, motivated by the AI agent point of view (e.g. of a robot discovering its environment autonomously), we consider how the same objective can discover the causal variables themselves, as a transformation of observed low-level variables with no causal meaning. Experiments in the two-variable case validate the proposed ideas and theoretical results.
****
Expected Information Maximization: Using the I-Projection for Mixture Density Estimation
Philipp Becker,Oleg Arenz,Gerhard Neumann
Modelling highly multi-modal data is a challenging problem in machine learning. Most algorithms are based on maximizing the likelihood, which corresponds to the M(oment)-projection of the data distribution to the model distribution.
The M-projection forces the model to average over modes it cannot represent. In contrast, the I(nformation)-projection ignores such modes in the data and concentrates on the modes the model can represent. Such behavior is appealing whenever we deal with highly multi-modal data where modelling single modes correctly is more important than covering all the modes. Despite this advantage, the I-projection is rarely used in practice due to the lack of algorithms that can efficiently optimize it based on data. In this work, we present a new algorithm called Expected Information Maximization (EIM) for computing the I-projection solely based on samples for general latent variable models, where we focus on Gaussian mixtures models and Gaussian mixtures of experts. Our approach applies a variational upper bound to the I-projection objective which decomposes the original objective into single objectives for each mixture component as well as for the coefficients, allowing an efficient optimization. Similar to GANs, our approach employs discriminators but uses a more stable optimization procedure, using a tight upper bound. We show that our algorithm is much more effective in computing the I-projection than recent GAN approaches and we illustrate the effectiveness of our approach for modelling multi-modal behavior on two pedestrian and traffic prediction datasets.
****
All Simulations Are Not Equal: Simulation Reweighing for Imperfect Information Games
Qucheng Gong,Yuandong Tian
Imperfect information games are challenging benchmarks for artificial intelligent systems. To reason and plan under uncertainty is a key towards general AI. Traditionally, large amounts of simulations are used in imperfect information games, and they sometimes perform sub-optimally due to large state and action spaces. In this work, we propose a simulation reweighing mechanism using neural networks. It performs backwards verification to public previous actions and assign proper belief weights to the simulations from the information set of the current observation, using an incomplete state solver network (ISSN). We use simulation rew

eighing in the playing phase of the game contract bridge, and show that it outpe
rforms previous state-of-the-art Monte Carlo simulation based methods, and achie
ves better play per decision.
****

Truth or backpropaganda? An empirical investigation of deep learning theory
Micah Goldblum,Jonas Geiping,Avi Schwarzschild,Michael Moeller,Tom Goldstein
We empirically evaluate common assumptions about neural networks that are widely
 held by practitioners and theorists alike. In this work, we: (1) prove the wid
espread existence of suboptimal local minima in the loss landscape of neural net
works, and we use our theory to find examples; (2) show that small-norm paramete
rs are not optimal for generalization; (3) demonstrate that ResNets do not confo
rm to wide-network theories, such as the neural tangent kernel, and that the int
eraction between skip connections and batch normalization plays a role; (4) fin
d that rank does not correlate with generalization or robustness in a practical
setting.
****

Learning to Rank Learning Curves
Martin Wistuba,Tejaswini Pedapati
Many automated machine learning methods, such as those for hyperparameter and ne
ural architecture optimization, are computationally expensive because they invol
ve training many different model configurations. In this work, we present a new
method that saves computational budget by terminating poor configurations early
on in the training. In contrast to existing methods, we consider this task as a
ranking and transfer learning problem. We qualitatively show that by optimizing
a pairwise ranking loss and leveraging learning curves from other data sets, our
 model is able to effectively rank learning curves without having to observe man
y or very long learning curves. We further demonstrate that our method can be us
ed to accelerate a neural architecture search by a factor of up to 100 without a
 significant performance degradation of the discovered architecture. In further
experiments we analyze the quality of ranking, the influence of different model
components as well as the predictive behavior of the model.
****

Set Functions for Time Series
Max Horn,Michael Moor,Christian Bock,Bastian Rieck,Karsten Borgwardt
Despite the eminent successes of deep neural networks, many architectures are of
ten hard to transfer to irregularly-sampled and asynchronous time series that oc
cur in many real-world datasets, such as healthcare applications. This paper pro
poses a novel framework for classifying irregularly sampled time series with una
ligned measurements, focusing on high scalability and data efficiency.
Our method SeFT (Set Functions for Time Series) is based on recent advances in d
ifferentiable set function learning, extremely parallelizable, and scales well t
o very large datasets and online monitoring scenarios.
We extensively compare our method to competitors on multiple healthcare time ser
ies datasets and show that it performs competitively whilst significantly reduci
ng runtime.
****

MissDeepCausal: causal inference from incomplete data using deep latent variable
 models
Julie Josse,Imke Mayer,Jean-Philippe Vert
Inferring causal effects of a treatment, intervention or policy from observation
al data is central to many applications. However, state-of-the-art methods for c
ausal inference seldom consider the possibility that covariates have missing val
ues, which is ubiquitous in many real-world analyses. Missing data greatly comp
licate causal inference procedures as they require an adapted unconfoundedness h
ypothesis which can be difficult to justify in practice. We circumvent this issu
e by considering latent confounders whose distribution is learned through variat
ional autoencoders adapted to missing values. They can be used either as a pre-p
rocessing step prior to causal inference but we also suggest to embed them in a
multiple imputation strategy to take into account the variability due to missing
 values. Numerical experiments demonstrate the effectiveness of the proposed me

thodology especially for non-linear models compared to competitors.
****
Variational Constrained Reinforcement Learning with Application to Planning at R oundabout
Yuan Tian,Minghao Han,Lixian Zhang,Wulong Liu,Jun Wang,Wei Pan
Planning at roundabout is crucial for autonomous driving in urban and rural envi ronments. Reinforcement learning is promising not only in dealing with complicat ed environment but also taking safety constraints into account as a as a constra ined Markov Decision Process. However, the safety constraints should be explicit ly mathematically formulated while this is challenging for planning at roundabou t due to unpredicted dynamic behavior of the obstacles. Therefore, to discrimina te the obstacles' states as either safe or unsafe is desired which is known as s ituation awareness modeling. In this paper, we combine variational learning and constrained reinforcement learning to simultaneously learn a Conditional Represe ntation Model (CRM) to encode the states into safe and unsafe distributions resp ectively as well as to learn the corresponding safe policy. Our approach is eva luated in using Simulation of Urban Mobility (SUMO) traffic simulator and it can generalize to various traffic flows.
****
Efficient Deep Representation Learning by Adaptive Latent Space Sampling
Yuanhan Mo,Shuo Wang,Chengliang Dai,Rui Zhou,Zhongzhao Teng,Wenjia Bai,Yike Guo
Supervised deep learning requires a large amount of training samples with annota tions (e.g. label class for classification task, pixel- or voxel-wised label map for segmentation tasks), which are expensive and time-consuming to obtain. Duri ng the training of a deep neural network, the annotated samples are fed into the network in a mini-batch way, where they are often regarded of equal importance. However, some of the samples may become less informative during training, as th e magnitude of the gradient start to vanish for these samples. In the meantime, other samples of higher utility or hardness may be more demanded for the trainin g process to proceed and require more exploitation. To address the challenges of expensive annotations and loss of sample informativeness, here we propose a nov el training framework which adaptively selects informative samples that are fed to the training process. The adaptive selection or sampling is performed based o n a hardness-aware strategy in the latent space constructed by a generative mode l. To evaluate the proposed training framework, we perform experiments on three different datasets, including MNIST and CIFAR-10 for image classification task a nd a medical image dataset IVUS for biophysical simulation task. On all three da tasets, the proposed framework outperforms a random sampling method, which demon strates the effectiveness of our framework.
****
Learning Functionally Decomposed Hierarchies for Continuous Navigation Tasks
Lukas Jendele,Sammy Christen,Emre Aksan,Otmar Hilliges
Solving long-horizon sequential decision making tasks in environments with spars e rewards is a longstanding problem in reinforcement learning (RL) research. Hie rarchical Reinforcement Learning (HRL) has held the promise to enhance the capab ilities of RL agents via operation on different levels of temporal abstraction. Despite the success of recent works in dealing with inherent nonstationarity and sample complexity, it remains difficult to generalize to unseen environments an d to transfer different layers of the policy to other agents. In this paper, we propose a novel HRL architecture, Hierarchical Decompositional Reinforcement Lea rning (HiDe), which allows decomposition of the hierarchical layers into indepen dent subtasks, yet allows for joint training of all layers in end-to-end manner. The main insight is to combine a control policy on a lower level with an image-based planning policy on a higher level. We evaluate our method on various compl ex continuous control tasks for navigation, demonstrating that generalization ac ross environments and transfer of higher level policies can be achieved. See vid eos https://sites.google.com/view/hide-rl
****
Deep Audio Priors Emerge From Harmonic Convolutional Networks
Zhoutong Zhang,Yunyun Wang,Chuang Gan,Jiajun Wu,Joshua B. Tenenbaum,Antonio Torr

alba,William T. Freeman
Convolutional neural networks (CNNs) excel in image recognition and generation. Among many efforts to explain their effectiveness, experiments show that CNNs carry strong inductive biases that capture natural image priors. Do deep networks also have inductive biases for audio signals? In this paper, we empirically show that current network architectures for audio processing do not show strong evidence in capturing such priors. We propose Harmonic Convolution, an operation that helps deep networks distill priors in audio signals by explicitly utilizing the harmonic structure within. This is done by engineering the kernel to be supported by sets of harmonic series, instead of local neighborhoods for convolutional kernels. We show that networks using Harmonic Convolution can reliably model audio priors and achieve high performance in unsupervised audio restoration tasks. With Harmonic Convolution, they also achieve better generalization performance for sound source separation.
****

Drawing Early-Bird Tickets: Toward More Efficient Training of Deep Networks
Haoran You,Chaojian Li,Pengfei Xu,Yonggan Fu,Yue Wang,Xiaohan Chen,Richard G. Baraniuk,Zhangyang Wang,Yingyan Lin
(Frankle & Carbin, 2019) shows that there exist winning tickets (small but critical subnetworks) for dense, randomly initialized networks, that can be trained alone to achieve comparable accuracies to the latter in a similar number of iterations. However, the identification of these winning tickets still requires the costly train-prune-retrain process, limiting their practical benefits. In this paper, we discover for the first time that the winning tickets can be identified at the very early training stage, which we term as Early-Bird (EB) tickets, via low-cost training schemes (e.g., early stopping and low-precision training) at large learning rates. Our finding of EB tickets is consistent with recently reported observations that the key connectivity patterns of neural networks emerge early. Furthermore, we propose a mask distance metric that can be used to identify EB tickets with low computational overhead, without needing to know the true winning tickets that emerge after the full training. Finally, we leverage the existence of EB tickets and the proposed mask distance to develop efficient training methods, which are achieved by first identifying EB tickets via low-cost schemes, and then continuing to train merely the EB tickets towards the target accuracy. Experiments based on various deep networks and datasets validate: 1) the existence of EB tickets and the effectiveness of mask distance in efficiently identifying them; and 2) that the proposed efficient training via EB tickets can achieve up to 5.8x ~ 10.7x energy savings while maintaining comparable or even better accuracy as compared to the most competitive state-of-the-art training methods, demonstrating a promising and easily adopted method for tackling cost-prohibitive deep network training.
****

On Understanding Knowledge Graph Representation
Carl Allen*,Ivana Balazevic*,Timothy M Hospedales
Many methods have been developed to represent knowledge graph data, which implicitly exploit low-rank latent structure in the data to encode known information and enable unknown facts to be inferred. To predict whether a relationship holds between entities, their embeddings are typically compared in the latent space following a relation-specific mapping. Whilst link prediction has steadily improved, the latent structure, and hence why such models capture semantic information, remains unexplained. We build on recent theoretical interpretation of word embeddings as a basis to consider an explicit structure for representations of relations between entities. For identifiable relation types, we are able to predict properties and justify the relative performance of leading knowledge graph representation methods, including their often overlooked ability to make independent predictions.
****

Encoding Musical Style with Transformer Autoencoders
Kristy Choi,Curtis Hawthorne,Ian Simon,Monica Dinculescu,Jesse Engel
We consider the problem of learning high-level controls over the global structur

e of sequence generation, particularly in the context of symbolic music generati on with complex language models. In this work, we present the Transformer autoen coder, which aggregates encodings of the input data across time to obtain a glob al representation of style from a given performance. We show it is possible to combine this global embedding with other temporally distributed embeddings, enab ling improved control over the separate aspects of performance style and and me lody. Empirically, we demonstrate the effectiveness of our method on a variety o f music generation tasks on the MAESTRO dataset and an internal, 10,000+ hour da taset of piano performances, where we achieve improvements in terms of log-likel ihood and mean listening scores as compared to relevant baselines.
****

Collaborative Inter-agent Knowledge Distillation for Reinforcement Learning
Zhang-Wei Hong,Prabhat Nagarajan,Guilherme Maeda
Reinforcement Learning (RL) has demonstrated promising results across several se quential decision-making tasks. However, reinforcement learning struggles to lea rn efficiently, thus limiting its pervasive application to several challenging p roblems. A typical RL agent learns solely from its own trial-and-error experienc es, requiring many experiences to learn a successful policy. To alleviate this p roblem, we propose collaborative inter-agent knowledge distillation (CIKD). CIKD is a learning framework that uses an ensemble of RL agents to execute different policies in the environment while sharing knowledge amongst agents in the ensem ble. Our experiments demonstrate that CIKD improves upon state-of-the-art RL met hods in sample efficiency and performance on several challenging MuJoCo benchmar k tasks. Additionally, we present an in-depth investigation on how CIKD leads to performance improvements.

****

Gauge Equivariant Spherical CNNs
Berkay Kicanaoglu,Pim de Haan,Taco Cohen
Spherical CNNs are convolutional neural networks that can process signals on the sphere, such as global climate and weather patterns or omnidirectional images. Over the last few years, a number of spherical convolution methods have been pro posed, based on generalized spherical FFTs, graph convolutions, and other ideas. However, none of these methods is simultaneously equivariant to 3D rotations, a ble to detect anisotropic patterns, computationally efficient, agnostic to the t ype of sample grid used, and able to deal with signals defined on only a part of the sphere. To address these limitations, we introduce the Gauge Equivariant Sp herical CNN. Our method is based on the recently proposed theory of Gauge Equiva riant CNNs, which is in principle applicable to signals on any manifold, and whi ch can be computed on any set of local charts covering all of the manifold or on ly part of it. In this paper we show how this method can be implemented efficien tly for the sphere, and show that the resulting method is fast, numerically accu rate, and achieves good results on the widely used benchmark problems of climate pattern segmentation and omnidirectional semantic segmentation.
****

Preventing Imitation Learning with Adversarial Policy Ensembles
Albert Zhan,Pieter Abbeel,Stas Tiomkin
Imitation learning can reproduce policies by observing experts, which poses a pr oblem regarding policy propriety. Policies, such as human, or policies on deploy ed robots, can all be cloned without consent from the owners. How can we protect our proprietary policies from cloning by an external observer? To answer this q uestion we introduce a new reinforcement learning framework, where we train an e nsemble of optimal policies, whose demonstrations are guaranteed to be useless f or an external observer. We formulate this idea by a constrained optimization pr oblem, where the objective is to improve proprietary policies, and at the same t ime deteriorate the virtual policy of an eventual external observer. We design a tractable algorithm to solve this new optimization problem by modifying the sta ndard policy gradient algorithm. It appears such problem formulation admits plau sible interpretations of confidentiality, adversarial behaviour, which enables a broader perspective of this work. We demonstrate explicitly the existence of su

ch 'non-clonable' ensembles, providing a solution to the above optimization prob lem, which is calculated by our modified policy gradient algorithm. To our knowl edge, this is the first work regarding the protection and privacy of policies in Reinforcement Learning.
****

Improving Generalization in Meta Reinforcement Learning using Learned Objectives
Louis Kirsch,Sjoerd van Steenkiste,Juergen Schmidhuber
Biological evolution has distilled the experiences of many learners into the gen eral learning algorithms of humans. Our novel meta reinforcement learning algori thm MetaGenRL is inspired by this process. MetaGenRL distills the experiences of many complex agents to meta-learn a low-complexity neural objective function th at decides how future individuals will learn. Unlike recent meta-RL algorithms, MetaGenRL can generalize to new environments that are entirely different from th ose used for meta-training. In some cases, it even outperforms human-engineered RL algorithms. MetaGenRL uses off-policy second-order gradients during meta-trai ning that greatly increase its sample efficiency.
****

A closer look at the approximation capabilities of neural networks
Kai Fong Ernest Chong
The universal approximation theorem, in one of its most general versions, says t hat if we consider only continuous activation functions σ, then a standard feedf orward neural network with one hidden layer is able to approximate any continuou s multivariate function f to any given approximation threshold ε, if and only if σ is non-polynomial. In this paper, we give a direct algebraic proof of the the orem. Furthermore we shall explicitly quantify the number of hidden units requir ed for approximation. Specifically, if X in R^n is compact, then a neural networ k with n input units, m output units, and a single hidden layer with {n+d choose d} hidden units (independent of m and ε), can uniformly approximate any polynom ial function f:X -> R^m whose total degree is at most d for each of its m coordi nate functions. In the general case that f is any continuous function, we show t here exists some N in O(ε^{-n}) (independent of m), such that N hidden units wou ld suffice to approximate f. We also show that this uniform approximation proper ty (UAP) still holds even under seemingly strong conditions imposed on the weigh ts. We highlight several consequences: (i) For any $\delta > 0$, the UAP still holds if we restrict all non-bias weights w in the last layer to satisfy $|w| < \delta$. (ii) T here exists some $\lambda>0$ (depending only on f and σ), such that the UAP still holds if we restrict all non-bias weights w in the first layer to satisfy $|w|>\lambda$. (iii) If the non-bias weights in the first layer are *fixed* and randomly chosen from a suitable range, then the UAP holds with probability 1.
****

VIMPNN: A physics informed neural network for estimating potential energies of o ut-of-equilibrium systems
Jay Morgan,Adeline Paiement,Christian Klinke
Simulation of molecular and crystal systems enables insight into interesting che mical properties that benefit processes ranging from drug discovery to material synthesis. However these simulations can be computationally expensive and time c onsuming despite the approximations through Density Functional Theory (DFT). We propose the Valence Interaction Message Passing Neural Network (VIMPNN) to appro ximate DFT's ground-state energy calculations. VIMPNN integrates physics prior k nowledge such as the existence of different interatomic bounds to estimate more accurate energies. Furthermore, while many previous machine learning methods con sider only stable systems, our proposed method is demonstrated on unstable syste ms at different atomic distances. VIMPNN predictions can be used to determine th e stable configurations of systems, i.e. stable distance for atoms -- a necessar y step for the future simulation of crystal growth for example. Our method is ex tensively evaluated on a augmented version of the QM9 dataset that includes unst able molecules, as well as a new dataset of infinite- and finite-size crystals, and is compared with the Message Passing Neural Network (MPNN). VIMPNN has compa rable accuracy with DFT, while allowing for 5 orders of magnitude in computation al speed up compared to DFT simulations, and produces more accurate and informat

ive potential energy curves than MPNN for estimating stable configurations.
****
SLM Lab: A Comprehensive Benchmark and Modular Software Framework for Reproducible Deep Reinforcement Learning
Wah Loon Keng,Laura Graesser,Milan Cvitkovic
We introduce SLM Lab, a software framework for reproducible reinforcement learning (RL) research. SLM Lab implements a number of popular RL algorithms, provides synchronous and asynchronous parallel experiment execution, hyperparameter search, and result analysis. RL algorithms in SLM Lab are implemented in a modular way such that differences in algorithm performance can be confidently ascribed to differences between algorithms, not between implementations. In this work we present the design choices behind SLM Lab and use it to produce a comprehensive single-codebase RL algorithm benchmark. In addition, as a consequence of SLM Lab's modular design, we introduce and evaluate a discrete-action variant of the Soft Actor-Critic algorithm (Haarnoja et al., 2018) and a hybrid synchronous/asynchronous training method for RL agents.
****
Data-Efficient Image Recognition with Contrastive Predictive Coding
Olivier J Henaff,Aravind Srinivas,Jeffrey De Fauw,Ali Razavi,Carl Doersch,S. M. Ali Eslami,Aaron van den Oord
Human observers can learn to recognize new categories of objects from a handful of examples, yet doing so with machine perception remains an open challenge. We hypothesize that data-efficient recognition is enabled by representations which make the variability in natural signals more predictable, as suggested by recent perceptual evidence. We therefore revisit and improve Contrastive Predictive Coding, a recently-proposed unsupervised learning framework, and arrive at a representation which enables generalization from small amounts of labeled data. When provided with only 1% of ImageNet labels (i.e. 13 per class), this model retains a strong classification performance, 73% Top-5 accuracy, outperforming supervised networks by 28% (a 65% relative improvement) and state-of-the-art semi-supervised methods by 14%. We also find this representation to serve as a useful substrate for object detection on the PASCAL-VOC 2007 dataset, approaching the performance of representations trained with a fully annotated ImageNet dataset.
****
Kaleidoscope: An Efficient, Learnable Representation For All Structured Linear Maps
Tri Dao,Nimit Sohoni,Albert Gu,Matthew Eichhorn,Amit Blonder,Megan Leszczynski,Atri Rudra,Christopher Ré
Modern neural network architectures use structured linear transformations, such as low-rank matrices, sparse matrices, permutations, and the Fourier transform, to improve inference speed and reduce memory usage compared to general linear maps. However, choosing which of the myriad structured transformations to use (and its associated parameterization) is a laborious task that requires trading off speed, space, and accuracy. We consider a different approach: we introduce a family of matrices called kaleidoscope matrices (K-matrices) that provably capture any structured matrix with near-optimal space (parameter) and time (arithmetic operation) complexity. We empirically validate that K-matrices can be automatically learned within end-to-end pipelines to replace hand-crafted procedures, in order to improve model quality. For example, replacing channel shuffles in ShuffleNet improves classification accuracy on ImageNet by up to 5%. K-matrices can also simplify hand-engineered pipelines---we replace filter bank feature computation in speech data preprocessing with a learnable kaleidoscope layer, resulting in only 0.4% loss in accuracy on the TIMIT speech recognition task. In addition, K-matrices can capture latent structure in models: for a challenging permuted image classification task, adding a K-matrix to a standard convolutional architecture can enable learning the latent permutation and improve accuracy by over 8 points. We provide a practically efficient implementation of our approach, and use K-matrices in a Transformer network to attain 36% faster end-to-end inference speed on a language translation task.
****

wMAN: WEAKLY-SUPERVISED MOMENT ALIGNMENT NETWORK FOR TEXT-BASED VIDEO SEGMENT RETRIEVAL

Reuben Tan,Huijuan Xu,Kate Saenko,Bryan A. Plummer

Given a video and a sentence, the goal of weakly-supervised video moment retrieval is to locate the video segment which is described by the sentence without having access to temporal annotations during training. Instead, a model must learn how to identify the correct segment (i.e. moment) when only being provided with video-sentence pairs. Thus, an inherent challenge is automatically inferring the latent correspondence between visual and language representations. To facilitate this alignment, we propose our Weakly-supervised Moment Alignment Network (wMAN) which exploits a multi-level co-attention mechanism to learn richer multimodal representations. The aforementioned mechanism is comprised of a Frame-By-Word interaction module as well as a novel Word-Conditioned Visual Graph (WCVG). Our approach also incorporates a novel application of positional encodings, commonly used in Transformers, to learn visual-semantic representations that contain contextual information of their relative positions in the temporal sequence through iterative message-passing. Comprehensive experiments on the DiDeMo and Charades-STA datasets demonstrate the effectiveness of our learned representations: our combined wMAN model not only outperforms the state-of-the-art weakly-supervised method by a significant margin but also does better than strongly-supervised state-of-the-art methods on some metrics.

****

Residual Energy-Based Models for Text Generation

Yuntian Deng,Anton Bakhtin,Myle Ott,Arthur Szlam,Marc'Aurelio Ranzato

Text generation is ubiquitous in many NLP tasks, from summarization, to dialogue and machine translation. The dominant parametric approach is based on locally normalized models which predict one word at a time. While these work remarkably well, they are plagued by exposure bias due to the greedy nature of the generation process. In this work, we investigate un-normalized energy-based models (EBMs) which operate not at the token but at the sequence level. In order to make training tractable, we first work in the residual of a pretrained locally normalized language model and second we train using noise contrastive estimation. Furthermore, since the EBM works at the sequence level, we can leverage pretrained bi-directional contextual representations, such as BERT and RoBERTa. Our experiments on two large language modeling datasets show that residual EBMs yield lower perplexity compared to locally normalized baselines. Moreover, generation via importance sampling is very efficient and of higher quality than the baseline models according to human evaluation.

****

AtomNAS: Fine-Grained End-to-End Neural Architecture Search

Jieru Mei,Yingwei Li,Xiaochen Lian,Xiaojie Jin,Linjie Yang,Alan Yuille,Jianchao Yang

Search space design is very critical to neural architecture search (NAS) algorithms. We propose a fine-grained search space comprised of atomic blocks, a minimal search unit that is much smaller than the ones used in recent NAS algorithms. This search space allows a mix of operations by composing different types of atomic blocks, while the search space in previous methods only allows homogeneous operations. Based on this search space, we propose a resource-aware architecture search framework which automatically assigns the computational resources (e.g., output channel numbers) for each operation by jointly considering the performance and the computational cost. In addition, to accelerate the search process, we propose a dynamic network shrinkage technique which prunes the atomic blocks with negligible influence on outputs on the fly. Instead of a search-and-retrain two-stage paradigm, our method simultaneously searches and trains the target architecture.

Our method achieves state-of-the-art performance under several FLOPs configurations on ImageNet with a small searching cost.

We open our entire codebase at: https://github.com/meijieru/AtomNAS.

****

AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty

Dan Hendrycks*,Norman Mu*,Ekin Dogus Cubuk,Barret Zoph,Justin Gilmer,Balaji Laks
hminarayanan
Modern deep neural networks can achieve high accuracy when the training distribu
tion and test distribution are identically distributed, but this assumption is f
requently violated in practice. When the train and test distributions are mismat
ched, accuracy can plummet. Currently there are few techniques that improve robu
stness to unforeseen data shifts encountered during deployment. In this work, we
 propose a technique to improve the robustness and uncertainty estimates of imag
e classifiers. We propose AugMix, a data processing technique that is simple to
implement, adds limited computational overhead, and helps models withstand unfor
eseen corruptions. AugMix significantly improves robustness and uncertainty meas
ures on challenging image classification benchmarks, closing the gap between pre
vious methods and the best possible performance in some cases by more than half.

****
Learning Latent Dynamics for Partially-Observed Chaotic Systems
Said ouala,Duong Nguyen,Lucas Drumetz,Bertrand Chapron,Ananda Pascual,Fabrice Co
llard,Lucile Gaultier,Ronan Fablet
This paper addresses the data-driven identification of latent representations of
 partially-observed dynamical systems, i.e. dynamical systems whose some compone
nts are never observed, with an emphasis on forecasting applications and long-te
rm asymptotic patterns. Whereas state-of-the-art data-driven approaches rely on
delay embeddings and linear decompositions of the underlying operators, we intro
duce a framework based on the data-driven identification of an augmented state-s
pace model using a neural-network-based representation. For a given training dat
aset, it amounts to jointly reconstructing the latent states and learning an ODE
 (Ordinary Differential Equation) representation in this space. Through numerica
l experiments, we demonstrate the relevance of the proposed framework w.r.t. sta
te-of-the-art approaches in terms of short-term forecasting errors and long-term
 behaviour. We further discuss how the proposed framework relates to Koopman ope
rator theory and Takens' embedding theorem.
****
Exploration via Flow-Based Intrinsic Rewards
Hsuan-Kung Yang,Po-Han Chiang,Min-Fong Hong,Chun-Yi Lee
Exploration bonuses derived from the novelty of observations in an environment h
ave become a popular approach to motivate exploration for reinforcement learning
 (RL) agents in the past few years. Recent methods such as curiosity-driven expl
oration usually estimate the novelty of new observations by the prediction error
s of their system dynamics models. In this paper, we introduce the concept of op
tical flow estimation from the field of computer vision to the RL domain and uti
lize the errors from optical flow estimation to evaluate the novelty of new obse
rvations. We introduce a flow-based intrinsic curiosity module (FICM) capable o
f learning the motion features and understanding the observations in a more comp
rehensive and efficient fashion. We evaluate our method and compare it with a nu
mber of baselines on several benchmark environments, including Atari games, Supe
r Mario Bros., and ViZDoom. Our results show that the proposed method is superio
r to the baselines in certain environments, especially for those featuring sophi
sticated moving patterns or with high-dimensional observation spaces.
****
Learning Underlying Physical Properties From Observations For Trajectory Predict
ion
Ekaterina Nikonova,Jochen Renz
In this work we present an approach that combines deep learning together with
laws of Newton’s physics for accurate trajectory predictions in physical games.
Our model learns to estimate physical properties and forces that generated given
observations, learns the relationships between available player’s actions and es
timated
physical properties and uses these extracted forces for predictions. We
show the advantages of using physical laws together with deep learning by evalua
ting

it against two baseline models that automatically discover features from the data without such a knowledge. We evaluate our model abilities to extract physical properties and to generalize to unseen trajectories in two games with a shooting mechanism. We also evaluate our model capabilities to transfer learned knowledge from a 2D game for predictions in a 3D game with a similar physics. We show that by using physical laws together with deep learning we achieve a bet ter
human-interpretability of learned physical properties, transfer of knowledge to a game with similar physics and very accurate predictions for previously unseen data.
****

SPREAD DIVERGENCE
Mingtian Zhang,David Barber,Thomas Bird,Peter Hayes,Raza Habib
For distributions p and q with different supports, the divergence $\div{p}{q }$ may not exist. We define a spread divergence $\sdiv{p}{q}$ on modified p an d q and describe sufficient conditions for the existence of such a divergence. We demonstrate how to maximize the discriminatory power of a given divergence b y parameterizing and learning the spread. We also give examples of using a sprea d divergence to train and improve implicit generative models, including linear m odels (Independent Components Analysis) and non-linear models (Deep Generative N etworks).
****

GraphQA: Protein Model Quality Assessment using Graph Convolutional Network
Federico Baldassarre,David Menéndez Hurtado,Arne Elofsson,Hossein Azizpour
Proteins are ubiquitous molecules whose function in biological processes is dete rmined by their 3D structure.
Experimental identification of a protein's structure can be time-consuming, proh ibitively expensive, and not always possible.
Alternatively, protein folding can be modeled using computational methods, which however are not guaranteed to always produce optimal results.
GraphQA is a graph-based method to estimate the quality of protein models, that possesses favorable properties such as representation learning, explicit modelin g of both sequential and 3D structure, geometric invariance and computational ef ficiency.
In this work, we demonstrate significant improvements of the state-of-the-art fo r both hand-engineered and representation-learning approaches, as well as carefu lly evaluating the individual contributions of GraphQA.
****

Disentanglement by Nonlinear ICA with General Incompressible-flow Networks (GIN)
Peter Sorrenson,Carsten Rother,Ullrich Köthe
A central question of representation learning asks under which conditions it is possible to reconstruct the true latent variables of an arbitrarily complex gene rative process. Recent breakthrough work by Khemakhem et al. (2019) on nonlinear ICA has answered this question for a broad class of conditional generative proc esses. We extend this important result in a direction relevant for application t o real-world data. First, we generalize the theory to the case of unknown intrin sic problem dimension and prove that in some special (but not very restrictive) cases, informative latent variables will be automatically separated from noise b y an estimating model. Furthermore, the recovered informative latent variables w ill be in one-to-one correspondence with the true latent variables of the genera ting process, up to a trivial component-wise transformation. Second, we introduc e a modification of the RealNVP invertible neural network architecture (Dinh et al. (2016)) which is particularly suitable for this type of problem: the General Incompressible-flow Network (GIN). Experiments on artificial data and EMNIST de monstrate that theoretical predictions are indeed verified in practice. In parti cular, we provide a detailed set of exactly 22 informative latent variables extr acted from EMNIST.
****

DEEP GRAPH SPECTRAL EVOLUTION NETWORKS FOR GRAPH TOPOLOGICAL TRANSFORMATION
Liang Zhao,Qingzhe Li,Negar Etemadyrad,Xiaojie Guo

Characterizing the underlying mechanism of graph topological evolution from a source graph to a target graph has attracted fast increasing attention in the deep graph learning domain. However, there lacks expressive and efficient that can handle global and local evolution patterns between source and target graphs. On the other hand, graph topological evolution has been investigated in the graph signal processing domain historically, but it involves intensive labors to manually determine suitable prescribed spectral models and prohibitive difficulty to fit their potential combinations and compositions. To address these challenges, this paper proposes the deep Graph Spectral Evolution Network (GSEN) for modeling the graph topology evolution problem by the composition of newly-developed generalized graph kernels. GSEN can effectively fit a wide range of existing graph kernels and their combinations and compositions with the theoretical guarantee and experimental verification. GSEN has outstanding efficiency in terms of time complexity ($O(n)$) and parameter complexity ($O(1)$), where n is the number of nodes of the graph. Extensive experiments on multiple synthetic and real-world datasets have demonstrated outstanding performance.
****

Angular Visual Hardness
Beidi Chen,Weiyang Liu,Animesh Garg,Zhiding Yu,Anshumali Shrivastava,Jan Kautz,Anima Anandkumar
The mechanisms behind human visual systems and convolutional neural networks (CNNs) are vastly different. Hence, it is expected that they have different notions of ambiguity or hardness. In this paper, we make a surprising discovery: there exists a (nearly) universal score function for CNNs whose correlation with human visual hardness is statistically significant. We term this function as angular visual hardness (AVH) and in a CNN, it is given by the normalized angular distance between a feature embedding and the classifier weights of the corresponding target category. We conduct an in-depth scientific study. We observe that CNN models with the highest accuracy also have the best AVH scores. This agrees with an earlier finding that state-of-art models tend to improve on classification of harder training examples. We find that AVH displays interesting dynamics during training: it quickly reaches a plateau even though the training loss keeps improving. This suggests the need for designing better loss functions that can target harder examples more effectively. Finally, we empirically show significant improvement in performance by using AVH as a measure of hardness in self-training tasks.

****

Deep Relational Factorization Machines
Hongchang Gao,Gang Wu,Ryan Rossi,Viswanathan Swaminathan,Heng Huang
Factorization Machines (FMs) is an important supervised learning approach due to its unique ability to capture feature interactions when dealing with high-dimensional sparse data. However, FMs assume each sample is independently observed and hence incapable of exploiting the interactions among samples. On the contrary, Graph Neural Networks (GNNs) has become increasingly popular due to its strength at capturing the dependencies among samples. But unfortunately, it cannot efficiently handle high-dimensional sparse data, which is quite common in modern machine learning tasks. In this work, to leverage their complementary advantages and yet overcome their issues, we proposed a novel approach, namely Deep Relational Factorization Machines, which can capture both the feature interaction and the sample interaction. In particular, we disclosed the relationship between the feature interaction and the graph, which opens a brand new avenue to deal with high-dimensional features. Finally, we demonstrate the effectiveness of the proposed approach with experiments on several real-world datasets.
****

Towards Scalable Imitation Learning for Multi-Agent Systems with Graph Neural Networks
Siyu Zhou,Chaitanya Rajasekhar,Mariano J. Phielipp,Heni Ben Amor
We propose an implementation of GNN that predicts and imitates the motion be- haviors from observed swarm trajectory data. The network’s ability to capture inte

raction dynamics in swarms is demonstrated through transfer learning. We finally discuss the inherent availability and challenges in the scalability of GNN, and proposed a method to improve it with layer-wise tuning and mixing of data enabled by padding.
****
On the Parameterization of Gaussian Mean Field Posteriors in Bayesian Neural Networks
Jakub ■wi■tkowski,Kevin Roth,Bastiaan S. Veeling,Linh Tran,Joshua V. Dillon,Jasper Snoek,Stephan Mandt,Tim Salimans,Rodolphe Jenatton,Sebastian Nowozin
Variational Bayesian Inference is a popular methodology for approximating posterior distributions in Bayesian neural networks. Recent work developing this class of methods has explored ever richer parameterizations of the approximate posterior in the hope of improving performance. In contrast, here we share a curious experimental finding that suggests instead restricting the variational distribution to a more compact parameterization. For a variety of deep Bayesian neural networks trained using Gaussian mean-field variational inference, we find that the posterior standard deviations consistently exhibits strong low-rank structure after convergence. This means that by decomposing these variational parameters into a low-rank factorization, we can make our variational approximation more compact without decreasing the models' performance. What's more, we find that such factorized parameterizations are easier to train since they improve the signal-to-noise ratio of stochastic gradient estimates of the variational lower bound, resulting in faster convergence.
****
Memory-Based Graph Networks
Amir Hosein Khasahmadi,Kaveh Hassani,Parsa Moradi,Leo Lee,Quaid Morris
Graph neural networks (GNNs) are a class of deep models that operate on data with arbitrary topology represented as graphs. We introduce an efficient memory layer for GNNs that can jointly learn node representations and coarsen the graph. We also introduce two new networks based on this layer: memory-based GNN (MemGNN) and graph memory network (GMN) that can learn hierarchical graph representations. The experimental results shows that the proposed models achieve state-of-the-art results in eight out of nine graph classification and regression benchmarks. We also show that the learned representations could correspond to chemical features in the molecule data.

****
GQ-Net: Training Quantization-Friendly Deep Networks
Rundong Li,Rui Fan
Network quantization is a model compression and acceleration technique that has become essential to neural network deployment. Most quantization methods per- form fine-tuning on a pretrained network, but this sometimes results in a large loss in accuracy compared to the original network. We introduce a new technique to train quantization-friendly networks, which can be directly converted to an accurate quantized network without the need for additional fine-tuning. Our technique allows quantizing the weights and activations of all network layers down to 4 bits, achieving high efficiency and facilitating deployment in practical settings. Com- pared to other fully quantized networks operating at 4 bits, we show substantial improvements in accuracy, for example 66.68% top-1 accuracy on ImageNet using ResNet-18, compared to the previous state-of-the-art accuracy of 61.52% Louizos et al. (2019) and a full precision reference accuracy of 69.76%. We performed a thorough set of experiments to test the efficacy of our method and also conducted ablation studies on different aspects of the method and techniques to improve training stability and accuracy. Our codebase and trained models are available on GitHub.
****
ExpandNets: Linear Over-parameterization to Train Compact Convolutional Networks
Shuxuan Guo,Jose M. Alvarez,Mathieu Salzmann
In this paper, we introduce a novel approach to training a given compact network. To this end, we build upon over-parameterization, which typically improves bot

h optimization and generalization in neural network training, while being unnece ssary at inference time. We propose to expand each linear layer of the compact n etwork into multiple linear layers, without adding any nonlinearity. As such, th e resulting expanded network can benefit from over-parameterization during train ing but can be compressed back to the compact one algebraically at inference. As evidenced by our experiments, this consistently outperforms training the compac t network from scratch and knowledge distillation using a teacher. In this conte xt, we introduce several expansion strategies, together with an initialization s cheme, and demonstrate the benefits of our ExpandNets on several tasks, includin g image classification, object detection, and semantic segmentation.
****

Variational Template Machine for Data-to-Text Generation
Rong Ye,Wenxian Shi,Hao Zhou,Zhongyu Wei,Lei Li
How to generate descriptions from structured data organized in tables? Existing approaches using neural encoder-decoder models often suffer from lacking diversi ty. We claim that an open set of templates is crucial for enriching the phrase c onstructions and realizing varied generations.Learning such templates is prohibi tive since it often requires a large paired <table,description>, which is seldom available. This paper explores the problem of automatically learning reusable " templates" from paired and non-paired data. We propose the variational template machine (VTM), a novel method to generate text descriptions from data tables. Ou r contributions include: a) we carefully devise a specific model architecture a nd losses to explicitly disentangle text template and semantic content informati on, in the latent spaces, and b) we utilize both small parallel data and large r aw text without aligned tables to enrich the template learning. Experiments on d atasets from a variety of different domains show that VTM is able to generate mo re diversely while keeping a good fluency and quality.
****

Phase Transitions for the Information Bottleneck in Representation Learning
Tailin Wu,Ian Fischer
In the Information Bottleneck (IB), when tuning the relative strength between co mpression and prediction terms, how do the two terms behave, and what's their re lationship with the dataset and the learned representation? In this paper, we se t out to answer these questions by studying multiple phase transitions in the IB objective: $IB_\beta[p(z|x)] = I(X; Z) - \beta I(Y; Z)$ defined on the encoding distributi on $p(z|x)$ for input X, target Y and representation Z, where sudden jumps of $dI(Y ; Z)/d\beta$ and prediction accuracy are observed with increasing β. We introduce a d efinition for IB phase transitions as a qualitative change of the IB loss landsc ape, and show that the transitions correspond to the onset of learning new class es. Using second-order calculus of variations, we derive a formula that provides a practical condition for IB phase transitions, and draw its connection with th e Fisher information matrix for parameterized models. We provide two perspective s to understand the formula, revealing that each IB phase transition is finding a component of maximum (nonlinear) correlation between X and Y orthogonal to the learned representation, in close analogy with canonical-correlation analysis (C CA) in linear settings. Based on the theory, we present an algorithm for discove ring phase transition points. Finally, we verify that our theory and algorithm a ccurately predict phase transitions in categorical datasets, predict the onset o f learning new classes and class difficulty in MNIST, and predict prominent phas e transitions in CIFAR10.

****

PopSGD: Decentralized Stochastic Gradient Descent in the Population Model
Giorgi Nadiradze,Amirmojtaba Sabour,Aditya Sharma,Ilia Markov,Vitaly Aksenov,Dan Alistarh.
The population model is a standard way to represent large-scale decentralized distributed systems, in which agents with limited computational power interact in randomly chosen pairs, in order to collectively solve global computational tasks. In contrast with synchronous gossip models, nodes are anonymous, lack a common notion of time, and have no control over their scheduling. In this paper,

we examine whether large-scale distributed optimization can be performed in this extremely restrictive setting.

We introduce and analyze a natural decentralized variant of stochastic gradient descent (SGD), called PopSGD, in which every node maintains a local parameter, and is able to compute stochastic gradients with respect to this parameter. Every pair-wise node interaction performs a stochastic gradient step at each agent, followed by averaging of the two models. We prove that, under standard assumptions, SGD can converge even in this extremely loose, decentralized setting, for both convex and non-convex objectives. Moreover, surprisingly, in the former case, the algorithm can achieve linear speedup in the number of nodes n. Our analysis leverages a new technical connection between decentralized SGD and randomized load balancing, which enables us to tightly bound the concentration of node parameters. We validate our analysis through experiments, showing that PopSGD can achieve convergence and speedup for large-scale distributed learning tasks in a supercomputing environment.
****
Symmetric-APL Activations: Training Insights and Robustness to Adversarial Attacks
Mohammadamin Tavakoli,Forest Agostinelli,Pierre Baldi
Deep neural networks with learnable activation functions have shown superior performance over deep neural networks with fixed activation functions for many different problems. The adaptability of learnable activation functions adds expressive power to the model which results in better performance. Here, we propose a new learnable activation function based on Adaptive Piecewise Linear units (APL), which 1) gives equal expressive power to both the positive and negative halves on the input space and 2) is able to approximate any zero-centered continuous non-linearity in a closed interval. We investigate how the shape of the Symmetric-APL function changes during training and perform ablation studies to gain insight into the reason behind these changes. We hypothesize that these activation functions go through two distinct stages: 1) adding gradient information and 2) adding expressive power. Finally, we show that the use of Symmetric-APL activations can significantly increase the robustness of deep neural networks to adversarial attacks. Our experiments on both black-box and open-box adversarial attacks show that commonly-used architectures, namely Lenet, Network-in-Network, and ResNet-18 can be up to 51% more resistant to adversarial fooling by only using the proposed activation functions instead of ReLUs.
****
Hidden incentives for self-induced distributional shift
David Scott Krueger,Tegan Maharaj,Shane Legg,Jan Leike
Decisions made by machine learning systems have increasing influence on the world. Yet it is common for machine learning algorithms to assume that no such influence exists. An example is the use of the i.i.d. assumption in online learning for applications such as content recommendation, where the (choice of) content displayed can change users' perceptions and preferences, or even drive them away, causing a shift in the distribution of users. Generally speaking, it is possible for an algorithm to change the distribution of its own inputs. We introduce the term self-induced distributional shift (SIDS) to describe this phenomenon. A large body of work in reinforcement learning and causal machine learning aims to deal with distributional shift caused by deploying learning systems previously trained offline. Our goal is similar, but distinct: we point out that changes to the learning algorithm, such as the introduction of meta-learning, can reveal hidden incentives for distributional shift (HIDS), and aim to diagnose and prevent problems associated with hidden incentives. We design a simple environment as a "unit test" for HIDS, as well as a content recommendation environment which allows us to disentangle different types of SIDS. We demonstrate the potential for HIDS to cause unexpected or undesirable behavior in these environments, and propose and test a mitigation strategy.
****
The divergences minimized by non-saturating GAN training

Matt Shannon
Interpreting generative adversarial network (GAN) training as approximate diverg
ence minimization has been
theoretically insightful, has spurred discussion, and has lead to theoretically
and practically interesting
extensions such as f-GANs and Wasserstein GANs. For both classic GANs and f-GANs
, there is an original variant of training and a "non-saturating" variant which
uses an alternative form of generator gradient. The original variant is theoreti
cally easier to study, but for GANs the alternative variant performs better in p
ractice. The non-saturating scheme is often regarded as a simple modification to
 deal with optimization issues, but we show that in fact the non-saturating sche
me for GANs is effectively optimizing a reverse KL-like f-divergence. We also de
velop a number of theoretical tools to help compare and classify f-divergences.
We hope these results may help to clarify some of the theoretical discussion sur
rounding the divergence minimization view of GAN training.
****
The Differentiable Cross-Entropy Method
Brandon Amos,Denis Yarats
We study the Cross-Entropy Method (CEM) for the non-convex optimization of a con
tinuous and parameterized objective function and introduce a differentiable vari
ant (DCEM) that enables us to differentiate the output of CEM with respect to th
e objective function's parameters. In the machine learning setting this brings C
EM inside of the end-to-end learning pipeline in cases this has otherwise been i
mpossible. We show applications in a synthetic energy-based structured predictio
n task and in non-convex continuous control. In the control setting we show on t
he simulated cheetah and walker tasks that we can embed their optimal action seq
uences with DCEM and then use policy optimization to fine-tune components of the
 controller as a step towards combining model-based and model-free RL.
****
Atomic Compression Networks
Jonas Falkner,Josif Grabocka,Lars Schmidt-Thieme
Compressed forms of deep neural networks are essential in deploying large-scale
computational models on resource-constrained devices. Contrary to analogous
domains where large-scale systems are build as a hierarchical repetition of smal
l-
scale units, the current practice in Machine Learning largely relies on models w
ith
non-repetitive components. In the spirit of molecular composition with repeating
atoms, we advance the state-of-the-art in model compression by proposing Atomic
Compression Networks (ACNs), a novel architecture that is constructed by recursi
ve
repetition of a small set of neurons. In other words, the same neurons with the
same weights are stochastically re-positioned in subsequent layers of the networ
k.
Empirical evidence suggests that ACNs achieve compression rates of up to three
orders of magnitudes compared to fine-tuned fully-connected neural networks (88×
to 1116× reduction) with only a fractional deterioration of classification accur
acy
(0.15% to 5.33%). Moreover our method can yield sub-linear model complexities
and permits learning deep ACNs with less parameters than a logistic regression
with no decline in classification accuracy.
****
Continual learning with hypernetworks
Johannes von Oswald,Christian Henning,Benjamin F. Grewe,João Sacramento
Artificial neural networks suffer from catastrophic forgetting when they are seq
uentially trained on multiple tasks. To overcome this problem, we present a nove
l approach based on task-conditioned hypernetworks, i.e., networks that generate
 the weights of a target model based on task identity. Continual learning (CL) i
s less difficult for this class of models thanks to a simple key feature: instea
d of recalling the input-output relations of all previously seen data, task-cond

itioned hypernetworks only require rehearsing task-specific weight realizations, which can be maintained in memory using a simple regularizer. Besides achieving state-of-the-art performance on standard CL benchmarks, additional experiments on long task sequences reveal that task-conditioned hypernetworks display a very large capacity to retain previous memories. Notably, such long memory lifetimes are achieved in a compressive regime, when the number of trainable hypernetwork weights is comparable or smaller than target network size. We provide insight into the structure of low-dimensional task embedding spaces (the input space of the hypernetwork) and show that task-conditioned hypernetworks demonstrate transfer learning. Finally, forward information transfer is further supported by empirical results on a challenging CL benchmark based on the CIFAR-10/100 image datasets.

****

Few-Shot Regression via Learning Sparsifying Basis Functions

Yi Loo,Yiluan Guo,Ngai-Man Cheung

Recent few-shot learning algorithms have enabled models to quickly adapt to new tasks based on only a few training samples. Previous few-shot learning works have mainly focused on classification and reinforcement learning. In this paper, we propose a few-shot meta-learning system that focuses exclusively on regression tasks. Our model is based on the idea that the degree of freedom of the unknown function can be significantly reduced if it is represented as a linear combination of a set of sparsifying basis functions. This enables a few labeled samples to approximate the function. We design a Basis Function Learner network to encode basis functions for a task distribution, and a Weights Generator network to generate the weight vector for a novel task. We show that our model outperforms the current state of the art meta-learning methods in various regression tasks.

****

Removing input features via a generative model to explain their attributions to classifier's decisions

Chirag Agarwal,Dan Schonfeld,Anh Nguyen

Interpretability methods often measure the contribution of an input feature to an image classifier's decisions by heuristically removing it via e.g. blurring, adding noise, or graying out, which often produce unrealistic, out-of-samples. Instead, we propose to integrate a generative inpainter into three representative attribution map methods as a mechanism for removing input features. Compared to the original counterparts, our methods (1) generate more plausible counterfactual samples under the true data generating process; (2) are more robust to hyperparameter settings; and (3) localize objects more accurately. Our findings were consistent across both ImageNet and Places365 datasets and two different pairs of classifiers and inpainters.

****

Top-down training for neural networks

Shucong Zhang,Cong-Thanh Do,Rama Doddipatla,Erfan Loweimi,Peter Bell,Steve Renals

Vanishing gradients pose a challenge when training deep neural networks, resulting in the top layers (closer to the output) in the network learning faster when compared with lower layers closer to the input. Interpreting the top layers as a classifier and the lower layers a feature extractor, one can hypothesize that unwanted network convergence may occur when the classifier has overfit with respect to the feature extractor. This can lead to the feature extractor being under-trained, possibly failing to learn much about the patterns in the input data. To address this we propose a good classifier hypothesis: given a fixed classifier that partitions the space well, the feature extractor can be further trained to fit that classifier and learn the data patterns well. This alleviates the problem of under-training the feature extractor and enables the network to learn patterns in the data with small partial derivatives. We verify this hypothesis empirically and propose a novel top-down training method. We train all layers jointly, obtaining a good classifier from the top layers, which are then frozen. Following re-initialization, we retrain the bottom layers with respect to the frozen classifier. Applying this approach to a set of speech recognition experiments u

sing the Wall Street Journal and noisy CHiME-4 datasets we observe substantial a
ccuracy gains. When combined with dropout, our method enables connectionist temp
oral classification (CTC) models to outperform joint CTC-attention models, which
 have more capacity and flexibility.
****
Demystifying Graph Neural Network Via Graph Filter Assessment
Yewen Wang,Ziniu Hu,Yusong Ye,Yizhou Sun
Graph Neural Networks (GNNs) have received tremendous attention recently due to
their power in handling graph data for different downstream tasks across differe
nt application domains. The key of GNN is its graph convolutional filters, and r
ecently various kinds of filters are designed. However, there still lacks in-dep
th analysis on (1) Whether there exists a best filter that can perform best on a
ll graph data; (2) Which graph properties will influence the optimal choice of g
raph filter; (3) How to design appropriate filter adaptive to the graph data. In
 this paper, we focus on addressing the above three questions. We first propose
a novel assessment tool to evaluate the effectiveness of graph convolutional fil
ters for a given graph. Using the assessment tool, we find out that there is no
single filter as a `silver bullet' that perform the best on all possible graphs.
 In addition, different graph structure properties will influence the optimal gr
aph convolutional filter's design choice. Based on these findings, we develop Ad
aptive Filter Graph Neural Network (AFGNN), a simple but powerful model that can
 adaptively learn task-specific filter. For a given graph, it leverages graph fi
lter assessment as regularization and learns to combine from a set of base filte
rs. Experiments on both synthetic and real-world benchmark datasets demonstrate
that our proposed model can indeed learn an appropriate filter and perform well
on graph tasks.
****
Towards Certified Defense for Unrestricted Adversarial Attacks
Shengjia Zhao,Yang Song,Stefano Ermon
Certified defenses against adversarial examples are very important in safety-cri
tical applications of machine learning. However, existing certified defense stra
tegies only safeguard against perturbation-based adversarial attacks, where the
attacker is only allowed to modify normal data points by adding small perturbati
ons. In this paper, we provide certified defenses under the more general threat
model of unrestricted adversarial attacks. We allow the attacker to generate arb
itrary inputs to fool the classifier, and assume the attacker knows everything e
xcept the classifiers' parameters and the training dataset used to learn it. Lac
k of knowledge about the classifiers parameters prevents an attacker from genera
ting adversarial examples successfully. Our defense draws inspiration from diffe
rential privacy, and is based on intentionally adding noise to the classifier's
outputs to limit the attacker's knowledge about the parameters. We prove concret
e bounds on the minimum number of queries required for any attacker to generate
a successful adversarial attack. For a simple linear classifiers we prove that t
he bound is asymptotically optimal up to a constant by exhibiting an attack algo
rithm that achieves this lower bound. We empirically show the success of our def
ense strategy against strong black box attack algorithms.
****
Permutation Equivariant Models for Compositional Generalization in Language
Jonathan Gordon,David Lopez-Paz,Marco Baroni,Diane Bouchacourt
Humans understand novel sentences by composing meanings and roles of core langua
ge components. In contrast, neural network models for natural language modeling
fail when such compositional generalization is required. The main contribution o
f this paper is to hypothesize that language compositionality is a form of group
-equivariance. Based on this hypothesis, we propose a set of tools for construct
ing equivariant sequence-to-sequence models. Throughout a variety of experiments
 on the SCAN tasks, we analyze the behavior of existing models under the lens of
 equivariance, and demonstrate that our equivariant architecture is able to achi
eve the type compositional generalization required in human language understandi
ng.
****

Training binary neural networks with real-to-binary convolutions
Brais Martinez,Jing Yang,Adrian Bulat,Georgios Tzimiropoulos
This paper shows how to train binary networks to within a few percent points (~3 -5%) of the full precision counterpart. We first show how to build a strong base line, which already achieves state-of-the-art accuracy, by combining recently pr oposed advances and carefully adjusting the optimization procedure. Secondly, we show that by attempting to minimize the discrepancy between the output of the b inary and the corresponding real-valued convolution, additional significant accu racy gains can be obtained. We materialize this idea in two complementary ways: (1) with a loss function, during training, by matching the spatial attention map s computed at the output of the binary and real-valued convolutions, and (2) in a data-driven manner, by using the real-valued activations, available during inf erence prior to the binarization process, for re-scaling the activations right a fter the binary convolution. Finally, we show that, when putting all of our impr ovements together, the proposed model beats the current state of the art by more than 5% top-1 accuracy on ImageNet and reduces the gap to its real-valued count erpart to less than 3% and 5% top-1 accuracy on CIFAR-100 and ImageNet respectiv ely when using a ResNet-18 architecture. Code available at https://github.com/br ais-martinez/real2binary
****

DO-AutoEncoder: Learning and Intervening Bivariate Causal Mechanisms in Images
Tianshuo Cong,Dan Peng,Furui Liu,Zhitang Chen
Some fundamental limitations of deep learning have been exposed such as lacking generalizability and being vunerable to adversarial attack. Instead, researchers realize that causation is much more stable than association relationship in dat a. In this paper, we propose a new framework called do-calculus AutoEncoder(DO-A E) for deep representation learning that fully capture bivariate causal relation ship in the images which allows us to intervene in images generation process. DO -AE consists of two key ingredients: causal relationship mining in images and in tervention-enabling deep causal structured representation learning. The goal her e is to learn deep representations that correspond to the concepts in the physic al world as well as their causal structure. To verify the proposed method, we cr eate a dataset named PHY2D, which contains abstract graphic description in accor dance with the laws of physics. Our experiments demonstrate our method is able t o correctly identify the bivariate causal relationship between concepts in image s and the representation learned enables a do-calculus manipulation to images, w hich generates artificial images that might possibly break the physical law depe nding on where we intervene the causal system.
****

StructBERT: Incorporating Language Structures into Pre-training for Deep Languag e Understanding
Wei Wang,Bin Bi,Ming Yan,Chen Wu,Jiangnan Xia,Zuyi Bao,Liwei Peng,Luo Si
Recently, the pre-trained language model, BERT (and its robustly optimized versi on RoBERTa), has attracted a lot of attention in natural language understanding (NLU), and achieved state-of-the-art accuracy in various NLU tasks, such as sent iment classification, natural language inference, semantic textual similarity an d question answering. Inspired by the linearization exploration work of Elman, w e extend BERT to a new model, StructBERT, by incorporating language structures i nto pre-training. Specifically, we pre-train StructBERT with two auxiliary tasks to make the most of the sequential order of words and sentences, which leverage language structures at the word and sentence levels, respectively. As a result, the new model is adapted to different levels of language understanding required by downstream tasks.

The StructBERT with structural pre-training gives surprisingly good empirical re sults on a variety of downstream tasks, including pushing the state-of-the-art o n the GLUE benchmark to 89.0 (outperforming all published models at the time of model submission), the F1 score on SQuAD v1.1 question answering to 93.0, the ac curacy on SNLI to 91.7.
****

Multichannel Generative Language Models
Harris Chan,Jamie Kiros,William Chan
A channel corresponds to a viewpoint or transformation of an underlying meaning. A pair of parallel sentences in English and French express the same underlying meaning but through two separate channels corresponding to their languages. In this work, we present Multichannel Generative Language Models (MGLM), which models the joint distribution over multiple channels, and all its decompositions using a single neural network. MGLM can be trained by feeding it k way parallel-data, bilingual data, or monolingual data across pre-determined channels. MGLM is capable of both conditional generation and unconditional sampling. For conditional generation, the model is given a fully observed channel, and generates the k-1 channels in parallel. In the case of machine translation, this is akin to giving it one source, and the model generates k-1 targets. MGLM can also do partial conditional sampling, where the channels are seeded with prespecified words, and the model is asked to infill the rest. Finally, we can sample from MGLM unconditionally over all k channels. Our experiments on the Multi30K dataset containing English, French, Czech, and German languages suggest that the multitask training with the joint objective leads to improvements in bilingual translations. We provide a quantitative analysis of the quality-diversity trade-offs for different variants of the multichannel model for conditional generation, and a measurement of self-consistency during unconditional generation. We provide qualitative examples for parallel greedy decoding across languages and sampling from the joint distribution of the 4 languages.
****
Smooth markets: A basic mechanism for organizing gradient-based learners
David Balduzzi,Wojciech M. Czarnecki,Tom Anthony,Ian Gemp,Edward Hughes,Joel Leibo,Georgios Piliouras,Thore Graepel
With the success of modern machine learning, it is becoming increasingly important to understand and control how learning algorithms interact. Unfortunately, negative results from game theory show there is little hope of understanding or controlling general n-player games. We therefore introduce smooth markets (SM-games), a class of n-player games with pairwise zero sum interactions. SM-games codify a common design pattern in machine learning that includes some GANs, adversarial training, and other recent algorithms. We show that SM-games are amenable to analysis and optimization using first-order methods.
****
Enhancing the Transformer with explicit relational encoding for math problem solving
Imanol Schlag,Paul Smolensky,Roland Fernandez,Nebojsa Jojic,Jürgen Schmidhuber,Jianfeng Gao
We incorporate Tensor-Product Representations within the Transformer in order to better support the explicit representation of relation structure.
Our Tensor-Product Transformer (TP-Transformer) sets a new state of the art on the recently-introduced Mathematics Dataset containing 56 categories of free-form math word-problems.
The essential component of the model is a novel attention mechanism, called TP-Attention, which explicitly encodes the relations between each Transformer cell and the other cells from which values have been retrieved by attention. TP-Attention goes beyond linear combination of retrieved values, strengthening representation-building and resolving ambiguities introduced by multiple layers of regular attention.
The TP-Transformer's attention maps give better insights into how it is capable of solving the Mathematics Dataset's challenging problems.
Pretrained models and code will be made available after publication.
****
Ergodic Inference: Accelerate Convergence by Optimisation
Yichuan Zhang,José Miguel Hernández-Lobato
Statistical inference methods are fundamentally important in machine learning. Most state-of-the-art inference algorithms are
variants of Markov chain Monte Carlo (MCMC) or variational inference (VI). Howev

er, both methods struggle with limitations in practice: MCMC methods can be comp utationally demanding; VI methods may have large bias.
In this work, we aim to improve upon MCMC and VI by a novel hybrid method based on the idea of reducing simulation bias of finite-length MCMC chains using gradi ent-based optimisation. The proposed method can generate low-biased samples by i ncreasing the length of MCMC simulation and optimising the MCMC hyper-parameters , which offers attractive balance between approximation bias and computational e fficiency. We show that our method produces promising results on popular benchma rks when compared to recent hybrid methods of MCMC and VI.
****

SemanticAdv: Generating Adversarial Examples via Attribute-Conditional Image Edi ting
Haonan Qiu,Chaowei Xiao,Lei Yang,Xinchen Yan,HongLak Lee,Bo Li
Deep neural networks (DNNs) have achieved great success in various applications due to their strong expressive power. However, recent studies have shown that DN Ns are vulnerable to adversarial examples which are manipulated instances target ing to mislead DNNs to make incorrect predictions. Currently, most such adversar ial examples try to guarantee “subtle perturbation" by limiting the Lp norm of t he perturbation. In this paper, we aim to explore the impact of semantic manipul ation on DNNs predictions by manipulating the semantic attributes of images and generate “unrestricted adversarial examples". Such semantic based perturbation i s more practical compared with the Lp bounded perturbation. In particular, we pr opose an algorithm SemanticAdv which leverages disentangled semantic factors to generate adversarial perturbation by altering controlled semantic attributes to fool the learner towards various “adversarial" targets. We conduct extensive exp eriments to show that the semantic based adversarial examples can not only fool different learning tasks such as face verification and landmark detection, but a lso achieve high targeted attack success rate against real-world black-box servi ces such as Azure face verification service based on transferability. To further demonstrate the applicability of SemanticAdv beyond face recognition domain, we also generate semantic perturbations on street-view images. Such adversarial ex amples with controlled semantic manipulation can shed light on further understan ding about vulnerabilities of DNNs as well as potential defensive approaches.
****

Uncertainty - sensitive learning and planning with ensembles
Piotr Mi■o■,■ukasz Kuci■ski,Konrad Czechowski,Piotr Kozakowski,Maciej Klimek
We propose a reinforcement learning framework for discrete environments in which an agent optimizes its behavior on two timescales. For the short one, it uses t ree search methods to perform tactical decisions. The long strategic level is ha ndled with an ensemble of value functions learned using TD-like backups. Combi ning these two techniques brings synergies. The planning module performs \textit {what-if} analysis allowing to avoid short-term pitfalls and boost backups of th e value function. Notably, our method performs well in environments with sparse rewards where standard $TD(1)$ backups fail. On the other hand, the value functi ons compensate for inherent short-sightedness of planning. Importantly, we use e nsembles to measure the epistemic uncertainty of value functions. This serves tw o purposes: a) it stabilizes planning, b) it guides exploration.

We evaluate our methods on discrete environments with sparse rewards: the Deep s ea chain environment, toy Montezuma's Revenge, and Sokoban. In all the cases, we obtain speed-up of learning and boost to the final performance.
****

Fair Resource Allocation in Federated Learning
Tian Li,Maziar Sanjabi,Ahmad Beirami,Virginia Smith
Federated learning involves training statistical models in massive, heterogeneou s networks. Naively minimizing an aggregate loss function in such a network may disproportionately advantage or disadvantage some of the devices. In this work, we propose q-Fair Federated Learning (q-FFL), a novel optimization objective ins pired by fair resource allocation in wireless networks that encourages a more fa ir (specifically, a more uniform) accuracy distribution across devices in federa

ted networks. To solve q-FFL, we devise a communication-efficient method, q-FedAvg, that is suited to federated networks. We validate both the effectiveness of q-FFL and the efficiency of q-FedAvg on a suite of federated datasets with both convex and non-convex models, and show that q-FFL (along with q-FedAvg) outperforms existing baselines in terms of the resulting fairness, flexibility, and efficiency.
****

Continual Learning via Principal Components Projection
Gyuhak Kim,Bing Liu

Continual learning in neural networks (NN) often suffers from catastrophic forgetting. That is, when learning a sequence of tasks on an NN, the learning of a new task will cause weight changes that may destroy the learned knowledge embedded in the weights for previous tasks. Without solving this problem, it is difficult to use an NN to perform continual or lifelong learning. Although researchers have attempted to solve the problem in many ways, it remains to be challenging. In this paper, we propose a new approach, called principal components projection (PCP). The idea is that in learning a new task, if we can ensure that the gradient updates will only occur in the orthogonal directions to the input vectors of the previous tasks, then the weight updates for learning the new task will not affect the previous tasks. We propose to compute the principal components of the input vectors and use them to transform the input and to project the gradient updates for learning each new task. PCP does not need to store any sampled data from previous tasks or to generate pseudo data of previous tasks and use them to help learn a new task. Empirical evaluation shows that the proposed method PCP markedly outperforms the state-of-the-art baseline methods.
****

Convolutional Conditional Neural Processes
Jonathan Gordon,Wessel P. Bruinsma,Andrew Y. K. Foong,James Requeima,Yann Dubois,Richard E. Turner

We introduce the Convolutional Conditional Neural Process (ConvCNP), a new member of the Neural Process family that models translation equivariance in the data. Translation equivariance is an important inductive bias for many learning problems including time series modelling, spatial data, and images. The model embeds data sets into an infinite-dimensional function space, as opposed to finite-dimensional vector spaces. To formalize this notion, we extend the theory of neural representations of sets to include functional representations, and demonstrate that any translation-equivariant embedding can be represented using a convolutional deep-set. We evaluate ConvCNPs in several settings, demonstrating that they achieve state-of-the-art performance compared to existing NPs. We demonstrate that building in translation equivariance enables zero-shot generalization to challenging, out-of-domain tasks.
****

Self-Induced Curriculum Learning in Neural Machine Translation
Dana Ruiter,Cristina España-Bonet,Josef van Genabith

Self-supervised neural machine translation (SS-NMT) learns how to extract/select suitable training data from comparable (rather than parallel) corpora and how to translate, in a way that the two tasks support each other in a virtuous circle. SS-NMT has been shown to be competitive with state-of-the-art unsupervised NMT. In this study we provide an in-depth analysis of the sampling choices the SS-NMT model takes during training. We show that, without it having been told to do so, the model selects samples of increasing (i) complexity and (ii) task-relevance in combination with (iii) a denoising curriculum. We observe that the dynamics of the mutual-supervision of both system internal representation types is vital for the extraction and hence translation performance. We show that in terms of the human Gunning-Fog Readability index (GF), SS-NMT starts by extracting and learning from Wikipedia data suitable for high school (GF=10--11) and quickly moves towards content suitable for first year undergraduate students (GF=13).
****

A Quality-Diversity Controllable GAN for Text Generation
Xingyu Lou,Kaihe Xu,Zhongliang Li,Tian Xia,Shaojun Wang,Jing Xiao

Text generation is a critical and difficult natural language processing task. Maximum likelihood estimate (MLE) based models have been arguably suffered from exposure bias in the inference stage and thus varieties of language generative adversarial networks (GANs) bypassing this problem have emerged. However, recent study has demonstrated that MLE models can constantly outperform GANs models over quality-diversity space under several metrics. In this paper, we propose a quality-diversity controllable language GAN.
****
Hydra: Preserving Ensemble Diversity for Model Distillation
Linh Tran,Bastiaan S. Veeling,Kevin Roth,Jakub ■wi■tkowski,Joshua V. Dillon,Jasper Snoek,Stephan Mandt,Tim Salimans,Sebastian Nowozin,Rodolphe Jenatton
Ensembles of models have been empirically shown to improve predictive performance and to yield robust measures of uncertainty. However, they are expensive in computation and memory. Therefore, recent research has focused on distilling ensembles into a single compact model, reducing the computational and memory burden of the ensemble while trying to preserve its predictive behavior. Most existing distillation formulations summarize the ensemble by capturing its average predictions. As a result, the diversity of the ensemble predictions, stemming from each individual member, is lost. Thus the distilled model cannot provide a measure of uncertainty comparable to that of the original ensemble. To retain more faithfully the diversity of the ensemble, we propose a distillation method based on a single multi-headed neural network, which we refer to as Hydra. The shared body network learns a joint feature representation that enables each head to capture the predictive behavior of each ensemble member. We demonstrate that with a slight increase in parameter count, Hydra improves distillation performance on classification and regression settings while capturing the uncertainty behaviour of the original ensemble over both in-domain and out-of-distribution tasks.
****
Few-Shot Few-Shot Learning and the role of Spatial Attention
Yann Lifchitz,Yannis Avrithis,Sylvaine Picard
Few-shot learning is often motivated by the ability of humans to learn new tasks from few examples. However, standard few-shot classification benchmarks assume that the representation is learned on a limited amount of base class data, ignoring the amount of prior knowledge that a human may have accumulated before learning new tasks. At the same time, even if a powerful representation is available, it may happen in some domain that base class data are limited or non-existent. This motivates us to study a problem where the representation is obtained from a classifier pre-trained on a large-scale dataset of a different domain, assuming no access to its training process, while the base class data are limited to few examples per class and their role is to adapt the representation to the domain at hand rather than learn from scratch. We adapt the representation in two stages, namely on the few base class data if available and on the even fewer data of new tasks. In doing so, we obtain from the pre-trained classifier a spatial attention map that allows focusing on objects and suppressing background clutter. This is important in the new problem, because when base class data are few, the network cannot learn where to focus implicitly. We also show that a pre-trained network may be easily adapted to novel classes, without meta-learning.
****
BAIL: Best-Action Imitation Learning for Batch Deep Reinforcement Learning
Xinyue Chen,Zijian Zhou,Zheng Wang,Che Wang,Yanqiu Wu,Qing Deng,Keith Ross
The field of Deep Reinforcement Learning (DRL) has recently seen a surge in research in batch reinforcement learning, which aims for sample-efficient learning from a given data set without additional interactions with the environment. In the batch DRL setting, commonly employed off-policy DRL algorithms can perform poorly and sometimes even fail to learn altogether. In this paper we propose anew algorithm, Best-Action Imitation Learning (BAIL), which unlike many off-policy DRL algorithms does not involve maximizing Q functions over the action space. Striving for simplicity as well as performance, BAIL first selects from the batch the actions it believes to be high-performing actions for their corresponding states; it then uses those state-action pairs to train a policy network using imita

tion learning. Although BAIL is simple, we demonstrate that BAIL achieves state of the art performance on the Mujoco benchmark, typically outperforming BatchConstrained deep Q-Learning (BCQ) by a wide margin.
****

Lossless Data Compression with Transformer
Gautier Izacard,Armand Joulin,Edouard Grave
Transformers have replaced long-short term memory and other recurrent neural networks variants in sequence modeling. It achieves state-of-the-art performance on a wide range of tasks related to natural language processing, including language modeling, machine translation, and sentence representation. Lossless compression is another problem that can benefit from better sequence models. It is closely related to the problem of online learning of language models. But, despite this ressemblance, it is an area where purely neural network based methods have not yet reached the compression ratio of state-of-the-art algorithms. In this paper, we propose a Transformer based lossless compression method that match the best compression ratio for text. Our approach is purely based on neural networks and does not rely on hand-crafted features as other lossless compression algorithms. We also provide a thorough study of the impact of the different components of the Transformer and its training on the compression ratio.
****

Meta-Learning with Warped Gradient Descent
Sebastian Flennerhag,Andrei A. Rusu,Razvan Pascanu,Francesco Visin,Hujun Yin,Raia Hadsell
Learning an efficient update rule from data that promotes rapid learning of new tasks from the same distribution remains an open problem in meta-learning. Typically, previous works have approached this issue either by attempting to train a neural network that directly produces updates or by attempting to learn better initialisations or scaling factors for a gradient-based update rule. Both of these approaches pose challenges. On one hand, directly producing an update forgoes a useful inductive bias and can easily lead to non-converging behaviour. On the other hand, approaches that try to control a gradient-based update rule typically resort to computing gradients through the learning process to obtain their meta-gradients, leading to methods that can not scale beyond few-shot task adaptation. In this work, we propose Warped Gradient Descent (WarpGrad), a method that intersects these approaches to mitigate their limitations. WarpGrad meta-learns an efficiently parameterised preconditioning matrix that facilitates gradient descent across the task distribution. Preconditioning arises by interleaving non-linear layers, referred to as warp-layers, between the layers of a task-learner. Warp-layers are meta-learned without backpropagating through the task training process in a manner similar to methods that learn to directly produce updates. WarpGrad is computationally efficient, easy to implement, and can scale to arbitrarily large meta-learning problems. We provide a geometrical interpretation of the approach and evaluate its effectiveness in a variety of settings, including few-shot, standard supervised, continual and reinforcement learning.
****

Never Give Up: Learning Directed Exploration Strategies
Adrià Puigdomènech Badia,Pablo Sprechmann,Alex Vitvitskyi,Daniel Guo,Bilal Piot,Steven Kapturowski,Olivier Tieleman,Martin Arjovsky,Alexander Pritzel,Andrew Bolt,Charles Blundell
We propose a reinforcement learning agent to solve hard exploration games by learning a range of directed exploratory policies. We construct an episodic memory-based intrinsic reward using k-nearest neighbors over the agent's recent experience to train the directed exploratory policies, thereby encouraging the agent to repeatedly revisit all states in its environment. A self-supervised inverse dynamics model is used to train the embeddings of the nearest neighbour lookup, biasing the novelty signal towards what the agent can control. We employ the framework of Universal Value Function Approximators to simultaneously learn many directed exploration policies with the same neural network, with different trade-offs between exploration and exploitation. By using the same neural network for different degrees of exploration/exploitation, transfer is demonstrated from predomi

nantly exploratory policies yielding effective exploitative policies. The propos ed method can be incorporated to run with modern distributed RL agents that coll ect large amounts of experience from many actors running in parallel on separate environment instances. Our method doubles the performance of the base agent in all hard exploration in the Atari-57 suite while maintaining a very high score a cross the remaining games, obtaining a median human normalised score of 1344.0%. Notably, the proposed method is the first algorithm to achieve non-zero rewards (with a mean score of 8,400) in the game of Pitfall! without using demonstratio ns or hand-crafted features.
****

AdvectiveNet: An Eulerian-Lagrangian Fluidic Reservoir for Point Cloud Processin g
Xingzhe He,Helen Lu Cao,Bo Zhu
This paper presents a novel physics-inspired deep learning approach for point cl oud processing motivated by the natural flow phenomena in fluid mechanics. Our learning architecture jointly defines data in an Eulerian world space, using a s tatic background grid, and a Lagrangian material space, using moving particles. By introducing this Eulerian-Lagrangian representation, we are able to naturally evolve and accumulate particle features using flow velocities generated from a generalized, high-dimensional force field. We demonstrate the efficacy of this system by solving various point cloud classification and segmentation problems w ith state-of-the-art performance. The entire geometric reservoir and data flow m imic the pipeline of the classic PIC/FLIP scheme in modeling natural flow, bridg ing the disciplines of geometric machine learning and physical simulation.
****

Unsupervised Spatiotemporal Data Inpainting
Yuan Yin,Arthur Pajot,Emmanuel de Bézenac,Patrick Gallinari
We tackle the problem of inpainting occluded area in spatiotemporal sequences, s uch as cloud occluded satellite observations, in an unsupervised manner. We plac e ourselves in the setting where there is neither access to paired nor unpaired training data. We consider several cases in which the underlying information of the observed sequence in certain areas is lost through an observation operator. In this case, the only available information is provided by the observation of t he sequence, the nature of the measurement process and its associated statistics . We propose an unsupervised-learning framework to retrieve the most probable se quence using a generative adversarial network. We demonstrate the capacity of ou r model to exhibit strong reconstruction capacity on several video datasets such as satellite sequences or natural videos.

****

Transferable Recognition-Aware Image Processing
Zhuang Liu,Tinghui Zhou,Zhiqiang Shen,Bingyi Kang,Trevor Darrell
Recent progress in image recognition has stimulated the deployment of vision sys tems (e.g. image search engines) at an unprecedented scale. As a result, visual data are now often consumed not only by humans but also by machines. Meanwhile, existing image processing methods only optimize for better human perception, whe reas the resulting images may not be accurately recognized by machines. This can be undesirable, e.g., the images can be improperly handled by search engines or recommendation systems. In this work, we propose simple approaches to improve m achine interpretability of processed images: optimizing the recognition loss dir ectly on the image processing network or through an intermediate transforming mo del, a process which we show can also be done in an unsupervised manner. Interes tingly, the processing model's ability to enhance the recognition performance ca n transfer when evaluated on different recognition models, even if they are of d ifferent architectures, trained on different object categories or even different recognition tasks. This makes the solutions applicable even when we do not have the knowledge about future downstream recognition models, e.g., if we are to up load the processed images to the Internet. We conduct comprehensive experiments on three image processing tasks with two downstream recognition tasks, and confi rm our method brings substantial accuracy improvement on both the same recogniti

on model and when transferring to a different one, with minimal or no loss in th
e image processing quality.
****

Transfer Active Learning For Graph Neural Networks
Shengding Hu,Meng Qu,Zhiyuan Liu,Jian Tang
Graph neural networks have been proved very effective for a variety of predictio
n tasks on graphs such as node classification. Generally, a large number of labe
led data are required to train these networks. However, in reality it could be v
ery expensive to obtain a large number of labeled data on large-scale graphs. In
 this paper, we studied active learning for graph neural networks, i.e., how to
effectively label the nodes on a graph for training graph neural networks. We fo
rmulated the problem as a sequential decision process, which sequentially label
informative nodes, and trained a policy network to maximize the performance of g
raph neural networks for a specific task. Moreover, we also studied how to learn
 a universal policy for labeling nodes on graphs with multiple training graphs a
nd then transfer the learned policy to unseen graphs. Experimental results on bo
th settings of a single graph and multiple training graphs (transfer learning se
tting) prove the effectiveness of our proposed approaches over many competitive
baselines.
****

Trajectory growth through random deep ReLU networks
Ilan Price,Jared Tanner
This paper considers the growth in the length of one-dimensional trajectories as
 they are passed through deep ReLU neural networks, which, among other things,
is one measure of the expressivity of deep networks. We generalise existing re
sults, providing an alternative, simpler method for lower bounding expected traj
ectory growth through random networks, for a more general class of weights distr
ibutions, including sparsely connected networks. We illustrate this approach by
 deriving bounds for sparse-Gaussian, sparse-uniform, and sparse-discrete-valued
 random nets. We prove that trajectory growth can remain exponential in depth wi
th these new distributions, including their sparse variants, with the sparsity p
arameter appearing in the base of the exponent.
****

Frequency Pooling: Shift-Equivalent and Anti-Aliasing Down Sampling
Zhendong Zhang,Cheolkon Jung
Convolutional layer utilizes the shift-equivalent prior of images which makes it
 a great success for image processing. However, commonly used down sampling meth
ods in convolutional neural networks (CNNs), such as max-pooling, average-poolin
g, and strided-convolution, are not shift-equivalent. This destroys the shift-eq
uivalent property of CNNs and degrades their performance. In this paper, we prop
ose a novel pooling method which is \emph{strict shift equivalent and anti-alias
ing} in theory. This is achieved by (inverse) Discrete Fourier Transform and we
call our method frequency pooling. Experiments on image classifications show tha
t frequency pooling improves accuracy and robustness w.r.t shifts of CNNs.
****

Improving Sequential Latent Variable Models with Autoregressive Flows
Joseph Marino,Lei Chen,Jiawei He,Stephan Mandt
We propose an approach for sequence modeling based on autoregressive normalizing
 flows. Each autoregressive transform, acting across time, serves as a moving re
ference frame for modeling higher-level dynamics. This technique provides a simp
le, general-purpose method for improving sequence modeling, with connections to
existing and classical techniques. We demonstrate the proposed approach both wit
h standalone models, as well as a part of larger sequential latent variable mode
ls. Results are presented on three benchmark video datasets, where flow-based dy
namics improve log-likelihood performance over baseline models.
****

SEED RL: Scalable and Efficient Deep-RL with Accelerated Central Inference
Lasse Espeholt,Raphaël Marinier,Piotr Stanczyk,Ke Wang,Marcin Michalski■
We present a modern scalable reinforcement learning agent called SEED (Scalable,
 Efficient Deep-RL). By effectively utilizing modern accelerators, we show that

it is not only possible to train on millions of frames per second but also to lo
wer the cost. of experiments compared to current methods. We achieve this with a
 simple architecture that features centralized inference and an optimized commun
ication layer. SEED adopts two state-of-the-art distributed algorithms, IMPALA/V
-trace (policy gradients) and R2D2 (Q-learning), and is evaluated on Atari-57, D
eepMind Lab and Google Research Football. We improve the state of the art on Foo
tball and are able to reach state of the art on Atari-57 twice as fast in wall-t
ime. For the scenarios we consider, a 40% to 80% cost reduction for running expe
riments is achieved. The implementation along with experiments is open-sourced s
o results can be reproduced and novel ideas tried out.
****

Sparse Transformer: Concentrated Attention Through Explicit Selection
Guangxiang Zhao,Junyang Lin,Zhiyuan Zhang,Xuancheng Ren,Xu Sun
Self-attention-based Transformer has demonstrated the state-of-the-art performan
ces in a number of natural language processing tasks. Self attention is able to
model long-term dependencies, but it may suffer from the extraction of irrelevan
t information in the context. To tackle the problem, we propose a novel model ca
lled Sparse Transformer. Sparse Transformer is able to improve the concentration
 of attention on the global context through an explicit selection of the most re
levant segments. Extensive experimental results on a series of natural language
processing tasks, including neural machine translation, image captioning, and la
nguage modeling, all demonstrate the advantages of Sparse Transformer in model p
erformance.
  Sparse Transformer reaches the state-of-the-art performances in the IWSLT 2015
 English-to-Vietnamese translation and IWSLT 2014 German-to-English translation.
 In addition, we conduct qualitative analysis to account for Sparse Transformer'
s superior performance.
****

Scheduled Intrinsic Drive: A Hierarchical Take on Intrinsically Motivated Explor
ation
Jingwei Zhang,Niklas Wetzel,Nicolai Dorka,Joschka Boedecker,Wolfram Burgard
Exploration in sparse reward reinforcement learning remains an open challenge. M
any state-of-the-art methods use intrinsic motivation to complement the sparse e
xtrinsic reward signal, giving the agent more opportunities to receive feedback
during exploration. Commonly these signals are added as bonus rewards, which res
ults in a mixture policy that neither conducts exploration nor task fulfillment
resolutely.
In this paper, we instead learn separate intrinsic and extrinsic task policies a
nd schedule between these different drives to accelerate exploration and stabili
ze learning. Moreover, we introduce a new type of intrinsic reward denoted as su
ccessor feature control (SFC), which is general and not task-specific. It takes
into account statistics over complete trajectories and thus differs from previou
s methods that only use local information to evaluate intrinsic motivation. We e
valuate our proposed scheduled intrinsic drive (SID) agent using three different
 environments with pure visual inputs: VizDoom, DeepMind Lab and DeepMind Contro
l Suite. The results show a substantially improved exploration efficiency with S
FC and the hierarchical usage of the intrinsic drives. A video of our experiment
al results can be found at https://gofile.io/?c=HpEwTd.
****

You CAN Teach an Old Dog New Tricks! On Training Knowledge Graph Embeddings
Daniel Ruffinelli,Samuel Broscheit,Rainer Gemulla
Knowledge graph embedding (KGE) models learn algebraic representations of the en
tities and relations in a knowledge graph. A vast number of KGE techniques for m
ulti-relational link prediction have been proposed in the recent literature, oft
en with state-of-the-art performance. These approaches differ along a number of
dimensions, including different model architectures, different training strategi
es, and different approaches to hyperparameter optimization. In this paper, we t
ake a step back and aim to summarize and quantify empirically the impact of each
 of these dimensions on model performance. We report on the results of an extens
ive experimental study with popular model architectures and training strategies

across a wide range of hyperparameter settings. We found that when trained appro priately, the relative performance differences between various model architectur es often shrinks and sometimes even reverses when compared to prior results. For example, RESCAL~\citep{nickel2011three}, one of the first KGE models, showed st rong performance when trained with state-of-the-art techniques; it was competiti ve to or outperformed more recent architectures. We also found that good (and of ten superior to prior studies) model configurations can be found by exploring re latively few random samples from a large hyperparameter space. Our results sugge st that many of the more advanced architectures and techniques proposed in the l iterature should be revisited to reassess their individual benefits. To foster f urther reproducible research, we provide all our implementations and experimenta l results as part of the open source LibKGE framework.
****

Unsupervised Learning of Graph Hierarchical Abstractions with Differentiable Coa rsening and Optimal Transport
Tengfei Ma,Jie Chen
Hierarchical abstractions are a methodology for solving large-scale graph proble ms in various disciplines. Coarsening is one such approach: it generates a pyram id of graphs whereby the one in the next level is a structural summary of the pr ior one. With a long history in scientific computing, many coarsening strategies were developed based on mathematically driven heuristics. Recently, resurgent i nterests exist in deep learning to design hierarchical methods learnable through differentiable parameterization. These approaches are paired with downstream ta sks for supervised learning. In this work, we propose an unsupervised approach, coined \textsc{OTCoarsening}, with the use of optimal transport. Both the coarse ning matrix and the transport cost matrix are parameterized, so that an optimal coarsening strategy can be learned and tailored for a given set of graphs. We de monstrate that the proposed approach produces meaningful coarse graphs and yield s competitive performance compared with supervised methods for graph classificat ion.
****

Defensive Tensorization: Randomized Tensor Parametrization for Robust Neural Net works
Adrian Bulat,Jean Kossaifi,Sourav Bhattacharya,Yannis Panagakis,Georgios Tzimiro poulos,Nicholas D. Lane,Maja Pantic
As deep neural networks become widely adopted for solving most problems in compu ter vision and audio-understanding, there are rising concerns about their potent ial vulnerability. In particular, they are very sensitive to adversarial attacks , which manipulate the input to alter models' predictions. Despite large bodies of work to address this issue, the problem remains open. In this paper, we prop ose defensive tensorization, a novel adversarial defense technique that leverage s a latent high order factorization of the network. Randomization is applied in the latent subspace, therefore resulting in dense reconstructed weights, without the sparsity or perturbations typically induced by the randomization.
Our approach can be easily integrated with any arbitrary neural architecture and combined with techniques like adversarial training. We empirically demonstrate the effectiveness of our approach on standard image classification benchmarks. W e further validate the generalizability of our approach across domains and low-p recision architectures by considering an audio classification task and binary ne tworks. In all cases, we demonstrate superior performance compared to prior work s in the target scenario.
****

Robust Reinforcement Learning via Adversarial Training with Langevin Dynamics
Huang Yu-Ting,Parameswaran Kamalaruban,Paul Rolland,Ya-Ping Hsieh,Volkan Cevher
We re-think the Two-Player Reinforcement Learning (RL) as an instance of a distr ibution sampling problem in infinite dimensions. Using the powerful Stochastic G radient Langevin Dynamics, we propose a new two-player RL algorithm, which is a sampling variant of the two-player policy gradient method. Our new algorithm con sistently outperforms existing baselines, in terms of generalization across diff ering training and testing conditions, on several MuJoCo environments.

****
Embodied Multimodal Multitask Learning
Devendra Singh Chaplot,Lisa Lee,Ruslan Salakhutdinov,Devi Parikh,Dhruv Batra
Visually-grounded embodied language learning models have recently shown to be effective at learning multiple multimodal tasks such as following navigational instructions and answering questions. In this paper, we address two key limitations of these models, (a) the inability to transfer the grounded knowledge across different tasks and (b) the inability to transfer to new words and concepts not seen during training using only a few examples. We propose a multitask model which facilitates knowledge transfer across tasks by disentangling the knowledge of words and visual attributes in the intermediate representations. We create scenarios and datasets to quantify cross-task knowledge transfer and show that the proposed model outperforms a range of baselines in simulated 3D environments. We also show that this disentanglement of representations makes our model modular and interpretable which allows for transfer to instructions containing new concepts.
****
High Fidelity Speech Synthesis with Adversarial Networks
Miko■aj Bi■kowski,Jeff Donahue,Sander Dieleman,Aidan Clark,Erich Elsen,Norman Casagrande,Luis C. Cobo,Karen Simonyan
Generative adversarial networks have seen rapid development in recent years and have led to remarkable improvements in generative modelling of images. However, their application in the audio domain has received limited attention, and autoregressive models, such as WaveNet, remain the state of the art in generative modelling of audio signals such as human speech. To address this paucity, we introduce GAN-TTS, a Generative Adversarial Network for Text-to-Speech. Our architecture is composed of a conditional feed-forward generator producing raw speech audio, and an ensemble of discriminators which operate on random windows of different sizes. The discriminators analyse the audio both in terms of general realism, as well as how well the audio corresponds to the utterance that should be pronounced. To measure the performance of GAN-TTS, we employ both subjective human evaluation (MOS - Mean Opinion Score), as well as novel quantitative metrics (Fréchet DeepSpeech Distance and Kernel DeepSpeech Distance), which we find to be well correlated with MOS. We show that GAN-TTS is capable of generating high-fidelity speech with naturalness comparable to the state-of-the-art models, and unlike autoregressive models, it is highly parallelisable thanks to an efficient feed-forward generator. Listen to GAN-TTS reading this abstract at https://storage.googleapis.com/deepmind-media/research/abstract.wav
****
Autoencoder-based Initialization for Recurrent Neural Networks with a Linear Memory
Antonio Carta,Alessandro Sperduti,Davide Bacciu
Orthogonal recurrent neural networks address the vanishing gradient problem by parameterizing the recurrent connections using an orthogonal matrix. This class of models is particularly effective to solve tasks that require the memorization of long sequences. We propose an alternative solution based on explicit memorization using linear autoencoders for sequences. We show how a recently proposed recurrent architecture, the Linear Memory Network, composed of a nonlinear feedforward layer and a separate linear recurrence, can be used to solve hard memorization tasks. We propose an initialization schema that sets the weights of a recurrent architecture to approximate a linear autoencoder of the input sequences, which can be found with a closed-form solution. The initialization schema can be easily adapted to any recurrent architecture.
    We argue that this approach is superior to a random orthogonal initialization due to the autoencoder, which allows the memorization of long sequences even before training. The empirical analysis show that our approach achieves competitive results against alternative orthogonal models, and the LSTM, on sequential MNIST, permuted MNIST and TIMIT.
****
Test-Time Training for Out-of-Distribution Generalization

Yu Sun,Xiaolong Wang,Zhuang Liu,John Miller,Alexei A. Efros,Moritz Hardt
We introduce a general approach, called test-time training, for improving the pe rformance of predictive models when test and training data come from different d istributions. Test-time training turns a single unlabeled test instance into a s elf-supervised learning problem, on which we update the model parameters before making a prediction on the test sample. We show that this simple idea leads to s urprising improvements on diverse image classification benchmarks aimed at evalu ating robustness to distribution shifts. Theoretical investigations on a convex model reveal helpful intuitions for when we can expect our approach to help.
****

Distance-based Composable Representations with Neural Networks
Graham Spinks,Marie-Francine Moens
We introduce a new deep learning technique that builds individual and class repr esentations based on distance estimates to randomly generated contextual dimensi ons for different modalities. Recent works have demonstrated advantages to creat ing representations from probability distributions over their contexts rather th an single points in a low-dimensional Euclidean vector space. These methods, how ever, rely on pre-existing features and are limited to textual information. In t his work, we obtain generic template representations that are vectors containing the average distance of a class to randomly generated contextual information. T hese representations have the benefit of being both interpretable and composable . They are initially learned by estimating the Wasserstein distance for differen t data subsets with deep neural networks. Individual samples or instances can th en be compared to the generic class representations, which we call templates, to determine their similarity and thus class membership. We show that this techniq ue, which we call WDVec, delivers good results for multi-label image classificat ion. Additionally, we illustrate the benefit of templates and their composabilit y by performing retrieval with complex queries where we modify the information c ontent in the representations. Our method can be used in conjunction with any ex isting neural network and create theoretically infinitely large feature maps.
****

At Stability's Edge: How to Adjust Hyperparameters to Preserve Minima Selection in Asynchronous Training of Neural Networks?
Niv Giladi,Mor Shpigel Nacson,Elad Hoffer,Daniel Soudry
Background: Recent developments have made it possible to accelerate neural netwo rks training significantly using large batch sizes and data parallelism. Trainin g in an asynchronous fashion, where delay occurs, can make training even more sc alable. However, asynchronous training has its pitfalls, mainly a degradation in generalization, even after convergence of the algorithm. This gap remains not w ell understood, as theoretical analysis so far mainly focused on the convergence rate of asynchronous methods.
Contributions: We examine asynchronous training from the perspective of dynamica l stability. We find that the degree of delay interacts with the learning rate, to change the set of minima accessible by an asynchronous stochastic gradient de scent algorithm. We derive closed-form rules on how the learning rate could be c hanged, while keeping the accessible set the same. Specifically, for high delay values, we find that the learning rate should be kept inversely proportional to the delay. We then extend this analysis to include momentum. We find momentum sh ould be either turned off, or modified to improve training stability. We provid e empirical experiments to validate our theoretical findings.
****

FRICATIVE PHONEME DETECTION WITH ZERO DELAY
Metehan Yurt,Alberto N. Escalante B.,Veniamin I. Morgenshtern
People with high-frequency hearing loss rely on hearing aids that employ frequen cy lowering algorithms. These algorithms shift some of the sounds from the high frequency band to the lower frequency band where the sounds become more percepti ble for the people with the condition. Fricative phonemes have an important part of their content concentrated in high frequency bands. It is important that the frequency lowering algorithm is activated exactly for the duration of a fricati ve phoneme, and kept off at all other times. Therefore, timely (with zero delay)

and accurate fricative phoneme detection is a key problem for high quality hearing aids. In this paper we present a deep learning based fricative phoneme detection algorithm that has zero detection delay and achieves state-of-the-art fricative phoneme detection accuracy on the TIMIT Speech Corpus. All reported results are reproducible and come with easy to use code that could serve as a baseline for future research.

****

Walking on the Edge: Fast, Low-Distortion Adversarial Examples
Hanwei Zhang,Teddy Furon,Yannis Avrithis,Laurent Amsaleg
Adversarial examples of deep neural networks are receiving ever increasing attention because they help in understanding and reducing the sensitivity to their input. This is natural given the increasing applications of deep neural networks in our everyday lives. When white-box attacks are almost always successful, it is typically only the distortion of the perturbations that matters in their evaluation.

In this work, we argue that speed is important as well, especially when considering that fast attacks are required by adversarial training. Given more time, iterative methods can always find better solutions. We investigate this speed-distortion trade-off in some depth and introduce a new attack called boundary projection BP that improves upon existing methods by a large margin. Our key idea is that the classification boundary is a manifold in the image space: we therefore quickly reach the boundary and then optimize distortion on this manifold.
****

Disentangling Trainability and Generalization in Deep Learning
Lechao Xiao,Jeffrey Pennington,Sam Schoenholz
A fundamental goal in deep learning is the characterization of trainability and generalization of neural networks as a function of their architecture and hyperparameters. In this paper, we discuss these challenging issues in the context of wide neural networks at large depths where we will see that the situation simplifies considerably. To do this, we leverage recent advances that have separately shown: (1) that in the wide network limit, random networks before training are Gaussian Processes governed by a kernel known as the Neural Network Gaussian Process (NNGP) kernel, (2) that at large depths the spectrum of the NNGP kernel simplifies considerably and becomes ``weakly data-dependent'', and (3) that gradient descent training of wide neural networks is described by a kernel called the Neural Tangent Kernel (NTK) that is related to the NNGP. Here we show that by combining the in the large depth limit the spectrum of the NTK simplifies in much the same way as that of the NNGP kernel. By analyzing this spectrum, we arrive at a precise characterization of trainability and generalization across a range of architectures including Fully Connected Networks (FCNs) and Convolutional Neural Networks (CNNs). We find that there are large regions of hyperparameter space where networks will train but will fail to generalize, in contrast with several recent results. By comparing CNNs with- and without-global average pooling, we show that CNNs without average pooling have very nearly identical learning dynamics to FCNs while CNNs with pooling contain a correction that alters its generalization performance. We perform a thorough empirical investigation of these theoretical results and finding excellent agreement on real datasets.
****

Provably Communication-efficient Data-parallel SGD via Nonuniform Quantization
Ali Ramezani-Kebrya,Fartash Faghri,Ilya Markov,Vitalii Aksenov,Dan Alistarh,Daniel M. Roy
As the size and complexity of models and datasets grow, so does the need for communication-efficient variants of stochastic gradient descent that can be deployed on clusters to perform model fitting in parallel. Alistarh et al. (2017) describe two variants of data-parallel SGD that quantize and encode gradients to lessen communication costs. For the first variant, QSGD, they provide strong theoretical guarantees. For the second variant, which we call QSGDinf, they demonstrate impressive empirical gains for distributed training of large neural networks. B

uilding on their work, we propose an alternative scheme for quantizing gradients and show that it yields stronger theoretical guarantees than exist for QSGD while matching the empirical performance of QSGDinf.
****

Functional Regularisation for Continual Learning with Gaussian Processes
Michalis K. Titsias,Jonathan Schwarz,Alexander G. de G. Matthews,Razvan Pascanu,Yee Whye Teh
We introduce a framework for Continual Learning (CL) based on Bayesian inference over the function space rather than the parameters of a deep neural network. This method, referred to as functional regularisation for Continual Learning, avoids forgetting a previous task by constructing and memorising an approximate posterior belief over the underlying task-specific function. To achieve this we rely on a Gaussian process obtained by treating the weights of the last layer of a neural network as random and Gaussian distributed. Then, the training algorithm sequentially encounters tasks and constructs posterior beliefs over the task-specific functions by using inducing point sparse Gaussian process methods. At each step a new task is first learnt and then a summary is constructed consisting of (i) inducing inputs – a fixed-size subset of the task inputs selected such that it optimally represents the task – and (ii) a posterior distribution over the function values at these inputs. This summary then regularises learning of future tasks, through Kullback-Leibler regularisation terms. Our method thus unites approaches focused on (pseudo-)rehearsal with those derived from a sequential Bayesian inference perspective in a principled way, leading to strong results on accepted benchmarks.
****

Verification of Generative-Model-Based Visual Transformations
Matthew Mirman,Timon Gehr,Martin Vechev
Generative networks are promising models for specifying visual transformations. Unfortunately, certification of generative models is challenging as one needs to capture sufficient non-convexity so to produce precise bounds on the output. Existing verification methods either fail to scale to generative networks or do not capture enough non-convexity. In this work, we present a new verifier, called ApproxLine, that can certify non-trivial properties of generative networks. ApproxLine performs both deterministic and probabilistic abstract interpretation and captures infinite sets of outputs of generative networks. We show that ApproxLine can verify interesting interpolations in the network's latent space.
****

A Graph Neural Network Assisted Monte Carlo Tree Search Approach to Traveling Salesman Problem
Zhihao Xing,Shikui Tu
We present a graph neural network assisted Monte Carlo Tree Search approach for the classical traveling salesman problem (TSP). We adopt a greedy algorithm framework to construct the optimal solution to TSP by adding the nodes successively. A graph neural network (GNN) is trained to capture the local and global graph structure and give the prior probability of selecting each vertex every step. The prior probability provides a heuristics for MCTS, and the MCTS output is an improved probability for selecting the successive vertex, as it is the feedback information by fusing the prior with the scouting procedure. Experimental results on TSP up to 100 nodes demonstrate that the proposed method obtains shorter tours than other learning-based methods.
****

Learning Likelihoods with Conditional Normalizing Flows
Christina Winkler,Daniel Worrall,Emiel Hoogeboom,Max Welling
Normalizing Flows (NFs) are able to model complicated distributions p(y) with strong inter-dimensional correlations and high multimodality by transforming a simple base density p(z) through an invertible neural network under the change of variables formula. Such behavior is desirable in multivariate structured prediction tasks, where handcrafted per-pixel loss-based methods inadequately capture strong correlations between output dimensions. We present a study of conditional normalizing flows (CNFs), a class of NFs where the base density to output space m

apping is conditioned on an input x, to model conditional densities p(y|x). CNFs are efficient in sampling and inference, they can be trained with a likelihood-based objective, and CNFs, being generative flows, do not suffer from mode collapse or training instabilities. We provide an effective method to train continuous CNFs for binary problems and in particular, we apply these CNFs to super-resolution and vessel segmentation tasks demonstrating competitive performance on standard benchmark datasets in terms of likelihood and conventional metrics.

****

Informed Temporal Modeling via Logical Specification of Factorial LSTMs

Hongyuan Mei,Guanghui Qin,Minjie Xu,Jason Eisner

Consider a world in which events occur that involve various entities. Learning how to predict future events from patterns of past events becomes more difficult as we consider more types of events. Many of the patterns detected in the dataset by an ordinary LSTM will be spurious since the number of potential pairwise correlations, for example, grows quadratically with the number of events. We propose a type of factorial LSTM architecture where different blocks of LSTM cells are responsible for capturing different aspects of the world state. We use Datalog rules to specify how to derive the LSTM structure from a database of facts about the entities in the world. This is analogous to how a probabilistic relational model (Getoor & Taskar, 2007) specifies a recipe for deriving a graphical model structure from a database. In both cases, the goal is to obtain useful inductive biases by encoding informed independence assumptions into the model. We specifically consider the neural Hawkes process, which uses an LSTM to modulate the rate of instantaneous events in continuous time. In both synthetic and real-world domains, we show that we obtain better generalization by using appropriate factorial designs specified by simple Datalog programs.

****

Regularly varying representation for sentence embedding

Hamid Jalalzai,Pierre Colombo,Chloé Clavel,Eric Gaussier,Giovanna Varni,Emmanuel Vignon,Anne Sabourin

The dominant approaches to sentence representation in natural language rely on learning embeddings on massive corpuses. The obtained embeddings have desirable properties such as compositionality and distance preservation (sentences with similar meanings have similar representations). In this paper, we develop a novel method for learning an embedding enjoying a dilation invariance property. We propose two algorithms: Orthrus, a classification algorithm, constrains the distribution of the embedded variable to be regularly varying, i.e. multivariate heavy-tail. and uses Extreme Value Theory (EVT) to tackle the classification task on two separate regions: the tail and the bulk. Hydra, a text generation algorithm for dataset augmentation, leverages the invariance property of the embedding learnt by Orthrus to generate coherent sentences with controllable attribute, e.g. positive or negative sentiment. Numerical experiments on synthetic and real text data demonstrate the relevance of the proposed framework.

****

A Simple and Scalable Shape Representation for 3D Reconstruction

Mateusz Michalkiewicz,Eugene Belilovsky,Mahsa Baktashmotagh,Anders Eriksson

Deep learning applied to the reconstruction of 3D shapes has seen growing interest. A popular approach to 3D reconstruction and generation in recent years has been the CNN decoder-encoder model often applied in voxel space. However this often scales very poorly with the resolution limiting the effectiveness of these models. Several sophisticated alternatives for decoding to 3D shapes have been proposed typically relying on alternative deep learning architectures. We show however in this work that standard benchmarks in 3D reconstruction can be tackled with a surprisingly simple approach: a linear decoder obtained by principal component analysis on the signed distance transform of the surface. This approach allows easily scaling to larger resolutions. We show in multiple experiments it is competitive with state of the art methods and also allows the decoder to be fine-tuned on the target task using a loss designed for SDF transforms, obtaining fur

ther gains.
****
Learning Through Limited Self-Supervision: Improving Time-Series Classification Without Additional Data via Auxiliary Tasks
Ian Fox,Harry Rubin-Falcone,Jenna Wiens
Self-supervision, in which a target task is improved without external supervision, has primarily been explored in settings that assume the availability of additional data. However, in many cases, particularly in healthcare, one may not have access to additional data (labeled or otherwise). In such settings, we hypothesize that self-supervision based solely on the structure of the data at-hand can help. We explore a novel self-supervision framework for time-series data, in which multiple auxiliary tasks (e.g., forecasting) are included to improve overall performance on a sequence-level target task without additional training data. We call this approach limited self-supervision, as we limit ourselves to only the data at-hand. We demonstrate the utility of limited self-supervision on three sequence-level classification tasks, two pertaining to real clinical data and one using synthetic data. Within this framework, we introduce novel forms of self-supervision and demonstrate their utility in improving performance on the target task. Our results indicate that limited self-supervision leads to a consistent improvement over a supervised baseline, across a range of domains. In particular, for the task of identifying atrial fibrillation from small amounts of electrocardiogram data, we observe a nearly 13% improvement in the area under the receiver operating characteristics curve (AUC-ROC) relative to the baseline (AUC-ROC=0.55 vs. AUC-ROC=0.62). Limited self-supervision applied to sequential data can aid in learning intermediate representations, making it particularly applicable in settings where data collection is difficult.
****
EvoNet: A Neural Network for Predicting the Evolution of Dynamic Graphs
Changmin Wu,Giannis Nikolentzos,Michalis Vazirgiannis
Neural networks for structured data like graphs have been studied extensively in recent years.
To date, the bulk of research activity has focused mainly on static graphs.
However, most real-world networks are dynamic since their topology tends to change over time.
Predicting the evolution of dynamic graphs is a task of high significance in the area of graph mining.
Despite its practical importance, the task has not been explored in depth so far, mainly due to its challenging nature.
In this paper, we propose a model that predicts the evolution of dynamic graphs.
Specifically, we use a graph neural network along with a recurrent architecture to capture the temporal evolution patterns of dynamic graphs.
Then, we employ a generative model which predicts the topology of the graph at the next time step and constructs a graph instance that corresponds to that topology.
We evaluate the proposed model on several artificial datasets following common network evolving dynamics, as well as on real-world datasets.
Results demonstrate the effectiveness of the proposed model.
****
Few-Shot One-Class Classification via Meta-Learning
Ahmed Frikha,Denis Krompaß,Hans-Georg Koepken,Volker Tresp
Although few-shot learning and one-class classification have been separately well studied, their intersection remains rather unexplored. Our work addresses the few-shot one-class classification problem and presents a meta-learning approach that requires only few data examples from only one class to adapt to unseen tasks. The proposed method builds upon the model-agnostic meta-learning (MAML) algorithm (Finn et al., 2017) and explicitly trains for few-shot class-imbalance learning, aiming to learn a model initialization that is particularly suited for learning one-class classification tasks after observing only a few examples of one class. Experimental results on datasets from the image domain and the time-series domain show that our model substantially outperforms the baselines, i

ncluding MAML, and demonstrate the ability to learn new tasks from only few majo
rity class samples. Moreover, we successfully learn anomaly detectors for a rea
l world application involving sensor readings recorded during industrial manufac
turing of workpieces with a CNC milling machine using only few examples from the
 normal class.
****
Training a Constrained Natural Media Painting Agent using Reinforcement Learning

Biao Jia,Jonathan Brandt,Radomir Mech,Ning Xu,Byungmoon Kim,Dinesh Manocha
We present a novel approach to train a natural media painting using reinforcemen
t learning. Given a reference image, our formulation is based on stroke-based r
endering that imitates human drawing and can be learned from scratch without su
pervision. Our painting agent computes a sequence of actions that represent the
primitive painting strokes. In order to ensure that the generated policy is pred
ictable and controllable, we use a constrained learning method and train the pai
nting agent using the environment model and follows the commands encoded in an
observation. We have applied our approach on many benchmarks and our results dem
onstrate that our constrained agent can handle different painting media and diff
erent constraints in the action space to collaborate with humans or other agents
.

****
The Role of Embedding Complexity in Domain-invariant Representations
Ching-Yao Chuang,Antonio Torralba,Stefanie Jegelka
Unsupervised domain adaptation aims to generalize the hypothesis trained in a so
urce domain to an unlabeled target domain. One popular approach to this problem
is to learn domain-invariant embeddings for both domains. In this work, we study
, theoretically and empirically, the effect of the embedding complexity on gener
alization to the target domain. In particular, this complexity affects an upper
bound on the target risk; this is reflected in experiments, too. Next, we specif
y our theoretical framework to multilayer neural networks. As a result, we devel
op a strategy that mitigates sensitivity to the embedding complexity, and empiri
cally achieves performance on par with or better than the best layer-dependent c
omplexity tradeoff.
****
Learning Curves for Deep Neural Networks: A field theory perspective
Omry Cohen,Or Malka,Zohar Ringel
A series of recent works established a rigorous correspondence between very wide
 deep neural networks (DNNs), trained in a particular manner, and noiseless Baye
sian Inference with a certain Gaussian Process (GP) known as the Neural Tangent
Kernel (NTK). Here we extend a known field-theory formalism for GP inference to
get a detailed understanding of learning-curves in DNNs trained in the regime of
 this correspondence (NTK regime). In particular, a renormalization-group approa
ch is used to show that noiseless GP inference using NTK, which lacks a good ana
lytical handle, can be well approximated by noisy GP inference on a related kern
el we call the renormalized NTK. Following this, a perturbation-theory analysis
is carried in one over the dataset-size yielding analytical expressions for the
(fixed-teacher/fixed-target) leading and sub-leading asymptotics of the learning
 curves. At least for uniform datasets, a coherent picture emerges wherein fully
-connected DNNs have a strong implicit bias towards functions which are low orde
r polynomials of the input.
****
Zero-Shot Policy Transfer with Disentangled Attention
Josh Roy,George Konidaris
Domain adaptation is an open problem in deep reinforcement learning (RL). Often,
 agents are asked to perform in environments where data is difficult to obtain.
In such settings, agents are trained in similar environments, such as simulators
, and are then transferred to the original environment. The gap between visual o
bservations of the source and target environments often causes the agent to fail
 in the target environment. We present a new RL agent, SADALA (Soft Attention Di

sentAngled representation Learning Agent). SADALA first learns a compressed state representation. It then jointly learns to ignore distracting features and solve the task presented. SADALA's separation of important and unimportant visual features leads to robust domain transfer. SADALA outperforms both prior disentangled-representation based RL and domain randomization approaches across RL environments (Visual Cartpole and DeepMind Lab).
****

Disentangled Cumulants Help Successor Representations Transfer to New Tasks
Chris Grimm,Irina Higgins,Andre Barreto,Denis Teplyashin,Markus Wulfmeier,Tim Hertweck,Raia Hadsell,Satinder Singh
Biological intelligence can learn to solve many diverse tasks in a data efficient manner by re-using basic knowledge and skills from one task to another. Furthermore, many of such skills are acquired through something called latent learning, where no explicit supervision for skill acquisition is provided. This is in contrast to the state-of-the-art reinforcement learning agents, which typically start learning each new task from scratch and struggle with knowledge transfer. In this paper we propose a principled way to learn and recombine a basis set of policies, which comes with certain guarantees on the coverage of the final task space. In particular, we construct a learning pipeline where an agent invests time to learn to perform intrinsically generated, goal-based tasks, and subsequently leverages this experience to quickly achieve a high level of performance on externally specified, often significantly more complex tasks through generalised policy improvement. We demonstrate both theoretically and empirically that such goal-based intrinsic tasks produce more transferable policies when the goals are specified in a space that exhibits a form of disentanglement.
****

Learning vector representation of local content and matrix representation of local motion, with implications for V1
Ruiqi Gao,Jianwen Xie,Siyuan Huang,Yufan Ren,Song-Chun Zhu,Ying Nian Wu
This paper proposes a representational model for image pair such as consecutive video frames that are related by local pixel displacements, in the hope that the model may shed light on motion perception in primary visual cortex (V1). The model couples the following two components. (1) The vector representations of local contents of images. (2) The matrix representations of local pixel displacements caused by the relative motions between the agent and the objects in the 3D scene. When the image frame undergoes changes due to local pixel displacements, the vectors are multiplied by the matrices that represent the local displacements. Our experiments show that our model can learn to infer local motions. Moreover, the model can learn Gabor-like filter pairs of quadrature phases.
****

Online Learned Continual Compression with Stacked Quantization Modules
Lucas Caccia,Eugene Belilovsky,Massimo Caccia,Joelle Pineau
We introduce and study the problem of Online Continual Compression, where one attempts to learn to compress and store a representative dataset from a non i.i.d data stream, while only observing each sample once. This problem is highly relevant for downstream online continual learning tasks, as well as standard learning methods under resource constrained data collection. We propose a new architecture which stacks Quantization Modules (SQM), consisting of a series of discrete autoencoders, each equipped with their own memory. Every added module is trained to reconstruct the latent space of the previous module using fewer bits, allowing the learned representation to become more compact as training progresses. This modularity has several advantages: 1) moderate compressions are quickly available early in training, which is crucial for remembering the early tasks, 2) as more data needs to be stored, earlier data becomes more compressed, freeing memory, 3) unlike previous methods, our approach does not require pretraining, even on challenging datasets. We show several potential applications of this method. We first replace the episodic memory used in Experience Replay with SQM, leading to significant gains on standard continual learning benchmarks using a fixed memory budget. We then apply our method to compressing larger images like those from Imagenet, and show that it is also effective with other modalities, such as LiD

AR data.
****
Gumbel-Matrix Routing for Flexible Multi-task Learning
Krzysztof Maziarz,Efi Kokiopoulou,Andrea Gesmundo,Luciano Sbaiz,Gabor Bartok,Jesse Berent
This paper proposes a novel per-task routing method for multi-task applications. Multi-task neural networks can learn to transfer knowledge across different tasks by using parameter sharing. However, sharing parameters between unrelated tasks can hurt performance. To address this issue, routing networks can be applied to learn to share each group of parameters with a different subset of tasks to better leverage tasks relatedness. However, this use of routing methods requires to address the challenge of learning the routing jointly with the parameters of a modular multi-task neural network. We propose the Gumbel-Matrix routing, a novel multi-task routing method based on the Gumbel-Softmax, that is designed to learn fine-grained parameter sharing. When applied to the Omniglot benchmark, the proposed method improves the state-of-the-art error rate by 17%.
****
The Frechet Distance of training and test distribution predicts the generalization gap
Julian Zilly,Hannes Zilly,Oliver Richter,Roger Wattenhofer,Andrea Censi,Emilio Frazzoli
Learning theory tells us that more data is better when minimizing the generalization error of identically distributed training and test sets. However, when training and test distribution differ, this distribution shift can have a significant effect. With a novel perspective on function transfer learning, we are able to lower bound the change of performance when transferring from training to test set with the Wasserstein distance between the embedded training and test set distribution. We find that there is a trade-off affecting performance between how invariant a function is to changes in training and test distribution and how large this shift in distribution is. Empirically across several data domains, we substantiate this viewpoint by showing that test performance correlates strongly with the distance in data distributions between training and test set. Complementary to the popular belief that more data is always better, our results highlight the utility of also choosing a training data distribution that is close to the test data distribution when the learned function is not invariant to such changes.
****
Selective sampling for accelerating training of deep neural networks
Berry Weinstein,Shai Fine,Yacov Hel-Or
We present a selective sampling method designed to accelerate the training of deep neural networks. To this end, we introduce a novel measurement, the {\it minimal margin score} (MMS), which measures the minimal amount of displacement an input should take until its predicted classification is switched. For multi-class linear classification, the MMS measure is a natural generalization of the margin-based selection criterion, which was thoroughly studied in the binary classification setting. In addition, the MMS measure provides an interesting insight into the progress of the training process and can be useful for designing and monitoring new training regimes. Empirically we demonstrate a substantial acceleration when training commonly used deep neural network architectures for popular image classification tasks. The efficiency of our method is compared against the standard training procedures, and against commonly used selective sampling alternatives: Hard negative mining selection, and Entropy-based selection.
Finally, we demonstrate an additional speedup when we adopt a more aggressive learning-drop regime while using the MMS selective sampling method.
****
Representing Unordered Data Using Multiset Automata and Complex Numbers
Justin DeBenedetto,David Chiang
Unordered, variable-sized inputs arise in many settings across multiple fields. The ability for set- and multiset- oriented neural networks to handle this type of input has been the focus of much work in recent years. We propose to represent multisets using complex-weighted multiset automata and show how the multiset

representations of certain existing neural architectures can be viewed as special cases of ours. Namely, (1) we provide a new theoretical and intuitive justification for the Transformer model's representation of positions using sinusoidal functions, and (2) we extend the DeepSets model to use complex numbers, enabling it to outperform the existing model on an extension of one of their tasks.

****
Robust Natural Language Representation Learning for Natural Language Inference by Projecting Superficial Words out
Wanyun Cui,Guangyu Zheng,Wei Wang
In natural language inference, the semantics of some words do not affect the inference. Such information is considered superficial and brings overfitting. How can we represent and discard such superficial information? In this paper, we use first order logic (FOL) - a classic technique from meaning representation language – to explain what information is superficial for a given sentence pair. Such explanation also suggests two inductive biases according to its properties. We proposed a neural network-based approach that utilizes the two inductive biases. We obtain substantial improvements over extensive experiments.
****
Deep Nonlinear Stochastic Optimal Control for Systems with Multiplicative Uncertainties
Marcus Pereira,Ziyi Wang,Tianrong Chen,Evangelos Theodorou
We present a deep recurrent neural network architecture to solve a class of stochastic optimal control problems described by fully nonlinear Hamilton Jacobi Bellman partial differential equations. Such PDEs arise when one considers stochastic dynamics characterized by uncertainties that are additive and control multiplicative. Stochastic models with the aforementioned characteristics have been used in computational neuroscience, biology, finance and aerospace systems and provide a more accurate representation of actuation than models with additive uncertainty. Previous literature has established the inadequacy of the linear HJB theory and instead rely on a non-linear Feynman-Kac lemma resulting in a second order forward-backward stochastic differential equations representation. However, the proposed solutions that use this representation suffer from compounding errors and computational complexity leading to lack of scalability. In this paper, we propose a deep learning based algorithm that leverages the second order Forward-Backward SDE representation and LSTM based recurrent neural networks to not only solve such Stochastic Optimal Control problems but also overcome the problems faced by previous approaches and scales well to high dimensional systems. The resulting control algorithm is tested on non-linear systems in robotics and biomechanics to demonstrate feasibility and out-performance against previous methods.

****
Compression based bound for non-compressed network: unified generalization error analysis of large compressible deep neural network
Taiji Suzuki,Hiroshi Abe,Tomoaki Nishimura
One of the biggest issues in deep learning theory is the generalization ability of networks with huge model size.
The classical learning theory suggests that overparameterized models cause overfitting.
However, practically used large deep models avoid overfitting, which is not well explained by the classical approaches.
To resolve this issue, several attempts have been made.
Among them, the compression based bound is one of the promising approaches.
However, the compression based bound can be applied only to a compressed network, and it is not applicable to the non-compressed original network.
In this paper, we give a unified frame-work that can convert compression based bounds to those for non-compressed original networks.
The bound gives even better rate than the one for the compressed network by improving the bias term.
By establishing the unified frame-work, we can obtain a data dependent generaliz

ation error bound which gives a tighter evaluation than the data independent one s.

****
Sentence embedding with contrastive multi-views learning
Antoine Simoulin
In this work, we propose a self-supervised method to learn sentence representations with an injection of linguistic knowledge. Multiple linguistic frameworks propose diverse sentence structures from which semantic meaning might be expressed out of compositional words operations. We aim to take advantage of this linguist diversity and learn to represent sentences by contrasting these diverse views. Formally, multiple views of the same sentence are mapped to close representations. On the contrary, views from other sentences are mapped further. By contrasting different linguistic views, we aim at building embeddings which better capture semantic and which are less sensitive to the sentence outward form.

****
Dynamics-Aware Embeddings
William Whitney,Rajat Agarwal,Kyunghyun Cho,Abhinav Gupta
In this paper we consider self-supervised representation learning to improve sample efficiency in reinforcement learning (RL). We propose a forward prediction objective for simultaneously learning embeddings of states and actions. These embeddings capture the structure of the environment's dynamics, enabling efficient policy learning. We demonstrate that our action embeddings alone improve the sample efficiency and peak performance of model-free RL on control from low-dimensional states. By combining state and action embeddings, we achieve efficient learning of high-quality policies on goal-conditioned continuous control from pixel observations in only 1-2 million environment steps.
****
AN ATTENTION-BASED DEEP NET FOR LEARNING TO RANK
Diego Klabjan,Baiyang Wang
In information retrieval, learning to rank constructs a machine-based ranking model which given a query, sorts the search results by their degree of relevance or importance to the query. Neural networks have been successfully applied to this problem, and in this paper, we propose an attention-based deep neural network which better incorporates different embeddings of the queries and search results with an attention-based mechanism. This model also applies a decoder mechanism to learn the ranks of the search results in a listwise fashion. The embeddings are trained with convolutional neural networks or the word2vec model. We demonstrate the performance of this model with image retrieval and text querying data sets.
****
RaPP: Novelty Detection with Reconstruction along Projection Pathway
Ki Hyun Kim,Sangwoo Shim,Yongsub Lim,Jongseob Jeon,Jeongwoo Choi,Byungchan Kim,Andre S. Yoon
We propose RaPP, a new methodology for novelty detection by utilizing hidden space activation values obtained from a deep autoencoder.
Precisely, RaPP compares input and its autoencoder reconstruction not only in the input space but also in the hidden spaces.
We show that if we feed a reconstructed input to the same autoencoder again, its activated values in a hidden space are equivalent to the corresponding reconstruction in that hidden space given the original input.
In order to aggregate the hidden space activation values, we propose two metrics, which enhance the novelty detection performance.
Through extensive experiments using diverse datasets, we validate that RaPP improves novelty detection performances of autoencoder-based approaches.
Besides, we show that RaPP outperforms recent novelty detection methods evaluated on popular benchmarks.

****

SAFE-DNN: A Deep Neural Network with Spike Assisted Feature Extraction for Noise Robust Inference
Xueyuan She,Priyabrata Saha,Daehyun Kim,Yun Long,Saibal Mukhopadhyay
We present a Deep Neural Network with Spike Assisted Feature Extraction (SAFE-DNN) to improve robustness of classification under stochastic perturbation of inputs. The proposed network augments a DNN with unsupervised learning of low-level features using spiking neuron network (SNN) with Spike-Time-Dependent-Plasticity (STDP). The complete network learns to ignore local perturbation while performing global feature detection and classification. The experimental results on CIFAR-10 and ImageNet subset demonstrate improved noise robustness for multiple DNN architectures without sacrificing accuracy on clean images.
****
Putting Machine Translation in Context with the Noisy Channel Model
Lei Yu,Laurent Sartran,Wojciech Stokowiec,Wang Ling,Lingpeng Kong,Phil Blunsom,Chris Dyer
We show that Bayes' rule provides a compelling mechanism for controlling unconditional document language models, using the long-standing challenge of effectively leveraging document context in machine translation. In our formulation, we estimate the probability of a candidate translation as the product of the unconditional probability of the candidate output document and the ``reverse translation probability'' of translating the candidate output back into the input source language document---the so-called ``noisy channel'' decomposition. A particular advantage of our model is that it requires only parallel sentences to train, rather than parallel documents, which are not always available. Using a new beam search reranking approximation to solve the decoding problem, we find that document language models outperform language models that assume independence between sentences, and that using either a document or sentence language model outperform comparable models that directly estimate the translation probability. We obtain the best-published results on the NIST Chinese--English translation task, a standard task for evaluating document translation. Our model also outperforms the benchmark Transformer model by approximately 2.5 BLEU on the WMT19 Chinese--English translation task.
****
Deep geometric matrix completion: Are we doing it right?
Amit Boyarski,Sanketh Vedula,Alex Bronstein
We address the problem of reconstructing a matrix from a subset of its entries. Current methods, branded as geometric matrix completion, augment classical rank regularization techniques by incorporating geometric information into the solution. This information is usually provided as graphs encoding relations between rows/columns.
In this work we propose a simple spectral approach for solving the matrix completion problem, via the framework of functional maps. We introduce the zoomout loss, a multiresolution spectral geometric loss inspired by recent advances in shape correspondence, whose minimization leads to state-of-the-art results on various recommender systems datasets. Surprisingly, for some datasets we were able to achieve comparable results even without incorporating geometric information. This puts into question both the quality of such information and current methods' ability to use it in a meaningful and efficient way.
****
Progressive Compressed Records: Taking a Byte Out of Deep Learning Data
Michael Kuchnik,George Amvrosiadis,Virginia Smith
Deep learning training accesses vast amounts of data at high velocity, posing challenges for datasets retrieved over commodity networks and storage devices. We introduce a way to dynamically reduce the overhead of fetching and transporting training data with a method we term Progressive Compressed Records (PCRs). PCRs deviate from previous formats by leveraging progressive compression to split each training example into multiple examples of increasingly higher fidelity, without adding to the total data size. Training examples of similar fidelity are grouped together, which reduces both the system overhead and data bandwidth needed to train a model. We show that models can be trained on aggressively compressed r

epresentations of the training data and still retain high accuracy, and that PCR
s can enable a 2x speedup on average over baseline formats using JPEG compressio
n. Our results hold across deep learning architectures for a wide range of datas
ets: ImageNet, HAM10000, Stanford Cars, and CelebA-HQ.
****
The Intriguing Effects of Focal Loss on the Calibration of Deep Neural Networks
Jishnu Mukhoti,Viveka Kulharia,Amartya Sanyal,Stuart Golodetz,Philip Torr,Puneet
 Dokania
Miscalibration -- a mismatch between a model's confidence and its correctness --
 of Deep Neural Networks (DNNs) makes their predictions hard for downstream comp
onents to trust. Ideally, we want networks to be accurate, calibrated and confid
ent. Temperature scaling, the most popular calibration approach, will calibrate
a DNN without affecting its accuracy, but it will also make its correct predicti
ons under-confident. In this paper, we show that replacing the widely used cross
-entropy loss with focal loss allows us to learn models that are already very we
ll calibrated. When combined with temperature scaling, focal loss, whilst preser
ving accuracy and yielding state-of-the-art calibrated models, also preserves th
e confidence of the model's correct predictions, which is extremely desirable fo
r downstream tasks. We provide a thorough analysis of the factors causing miscal
ibration, and use the insights we glean from this to theoretically justify the e
mpirically excellent performance of focal loss. We perform extensive experiments
 on a variety of computer vision (CIFAR-10/100) and NLP (SST, 20 Newsgroup) data
sets, and with a wide variety of different network architectures, and show that
our approach achieves state-of-the-art accuracy and calibration in almost all ca
ses.
****
Hypermodels for Exploration
Vikranth Dwaracherla,Xiuyuan Lu,Morteza Ibrahimi,Ian Osband,Zheng Wen,Benjamin V
an Roy
We study the use of hypermodels to represent epistemic uncertainty and guide exp
loration.
This generalizes and extends the use of ensembles to approximate Thompson sampli
ng. The computational cost of training an ensemble grows with its size, and as s
uch, prior work has typically been limited to ensembles with tens of elements. W
e show that alternative hypermodels can enjoy dramatic efficiency gains, enablin
g behavior that would otherwise require hundreds or thousands of elements, and e
ven succeed in situations where ensemble methods fail to learn regardless of siz
e.
This allows more accurate approximation of Thompson sampling as well as use of m
ore sophisticated exploration schemes. In particular, we consider an approximat
e form of information-directed sampling and demonstrate performance gains relati
ve to Thompson sampling. As alternatives to ensembles, we consider linear and n
eural network hypermodels, also known as hypernetworks.
We prove that, with neural network base models, a linear hypermodel can represen
t essentially any distribution over functions, and as such, hypernetworks do not
 extend what can be represented.
****
Denoising Improves Latent Space Geometry in Text Autoencoders
Tianxiao Shen,Jonas Mueller,Regina Barzilay,Tommi Jaakkola
Neural language models have recently shown impressive gains in unconditional tex
t generation, but controllable generation and manipulation of text remain challe
nging. In particular, controlling text via latent space operations in autoencode
rs has been difficult, in part due to chaotic latent space geometry. We propose
to employ adversarial autoencoders together with denoising (referred as DAAE) to
 drive the latent space to organize itself. Theoretically, we prove that input s
entence perturbations in the denoising approach encourage similar sentences to m
ap to similar latent representations. Empirically, we illustrate the trade-off b
etween text-generation and autoencoder-reconstruction capabilities, and our mode
l significantly improves over other autoencoder variants. Even from completely u
nsupervised training, DAAE can successfully alter the tense/sentiment of sentenc

es via simple latent vector arithmetic.
****
On Symmetry and Initialization for Neural Networks
Ido Nachum,Amir Yehudayoff
This work provides an additional step in the theoretical understanding of neural networks. We consider neural networks with one hidden layer and show that when learning symmetric functions, one can choose initial conditions so that standard SGD training efficiently produces generalization guarantees. We empirically verify this and show that this does not hold when the initial conditions are chosen at random. The proof of convergence investigates the interaction between the two layers of the network. Our results highlight the importance of using symmetry in the design of neural networks.
****
Meta Reinforcement Learning with Autonomous Inference of Subtask Dependencies
Sungryull Sohn,Hyunjae Woo,Jongwook Choi,Honglak Lee
We propose and address a novel few-shot RL problem, where a task is characterized by a subtask graph which describes a set of subtasks and their dependencies that are unknown to the agent. The agent needs to quickly adapt to the task over few episodes during adaptation phase to maximize the return in the test phase. Instead of directly learning a meta-policy, we develop a Meta-learner with Subtask Graph Inference (MSGI), which infers the latent parameter of the task by interacting with the environment and maximizes the return given the latent parameter. To facilitate learning, we adopt an intrinsic reward inspired by upper confidence bound (UCB) that encourages efficient exploration. Our experiment results on two grid-world domains and StarCraft II environments show that the proposed method is able to accurately infer the latent task parameter, and to adapt more efficiently than existing meta RL and hierarchical RL methods.
****
Policy path programming
Daniel McNamee
We develop a normative theory of hierarchical model-based policy optimization for Markov decision processes resulting in a full-depth, full-width policy iteration algorithm. This method performs policy updates which integrate reward information over all states at all horizons simultaneously thus sequentially maximizing the expected reward obtained per algorithmic iteration. Effectively, policy path programming ascends the expected cumulative reward gradient in the space of policies defined over all state-space paths. An exact formula is derived which finitely parametrizes these path gradients in terms of action preferences. Policy path gradients can be directly computed using an internal model thus obviating the need to sample paths in order to optimize in depth. They are quadratic in successor representation entries and afford natural generalizations to higher-order gradient techniques. In simulations, it is shown that intuitive hierarchical reasoning is emergent within the associated policy optimization dynamics.
****
Meta-Learning with Network Pruning for Overfitting Reduction
Hongduan Tian,Bo Liu,Xiao-Tong Yuan,Qingshan Liu
Meta-Learning has achieved great success in few-shot learning. However, the existing meta-learning models have been evidenced to overfit on meta-training tasks when using deeper and wider convolutional neural networks. This means that we cannot improve the meta-generalization performance by merely deepening or widening the networks. To remedy such a deficiency of meta-overfitting, we propose in this paper a sparsity constrained meta-learning approach to learn from meta-training tasks a subnetwork from which first-order optimization methods can quickly converge towards the optimal network in meta-testing tasks. Our theoretical analysis shows the benefit of sparsity for improving the generalization gap of the learned meta-initialization network. We have implemented our approach on top of the widely applied Reptile algorithm assembled with varying network pruning routines including Dense-Sparse-Dense (DSD) and Iterative Hard Thresholding (IHT). Extensive experimental results on benchmark datasets with different over-parameterized deep networks demonstrate that our method can not only effectively ease meta-

overfitting but also in many cases improve the meta-generalization performance w
hen applied to few-shot classification tasks.
****

Kernel and Rich Regimes in Overparametrized Models
Blake Woodworth,Suriya Gunasekar,Pedro Savarese,Edward Moroshko,Itay Golan,Jason
 Lee,Daniel Soudry,Nathan Srebro
A recent line of work studies overparametrized neural networks in the "kernel re
gime," i.e. when the network behaves during training as a kernelized linear pred
ictor, and thus training with gradient descent has the effect of finding the min
imum RKHS norm solution. This stands in contrast to other studies which demonst
rate how gradient descent on overparametrized multilayer networks can induce ric
h implicit biases that are not RKHS norms. Building on an observation by Chizat
 and Bach, we show how the scale of the initialization controls the transition b
etween the "kernel" (aka lazy) and "rich" (aka active) regimes and affects gener
alization properties in multilayer homogeneous models. We provide a complete an
d detailed analysis for a simple two-layer model that already exhibits an intere
sting and meaningful transition between the kernel and rich regimes, and we demo
nstrate the transition for more complex matrix factorization models and multilay
er non-linear networks.
****

A Boolean Task Algebra for Reinforcement Learning
Geraud Nangue Tasse,Steven James,Benjamin Rosman
We propose a framework for defining a Boolean algebra over the space of tasks. T
his allows us to formulate new tasks in terms of the negation, disjunction and c
onjunction of a set of base tasks. We then show that by learning goal-oriented v
alue functions and restricting the transition dynamics of the tasks, an agent ca
n solve these new tasks with no further learning. We prove that by composing the
se value functions in specific ways, we immediately recover the optimal policies
 for all tasks expressible under the Boolean algebra. We verify our approach in
two domains, including a high-dimensional video game environment requiring funct
ion approximation, where an agent first learns a set of base skills, and then co
mposes them to solve a super-exponential number of new tasks.
****

Explanation by Progressive Exaggeration
Sumedha Singla,Brian Pollack,Junxiang Chen,Kayhan Batmanghelich
As machine learning methods see greater adoption and implementation in high stak
es applications such as medical image diagnosis, the need for model interpretabi
lity and explanation has become more critical. Classical approaches that assess
feature importance (eg saliency maps) do not explain how and why a particular re
gion of an image is relevant to the prediction. We propose a method that explain
s the outcome of a classification black-box by gradually exaggerating the semant
ic effect of a given class. Given a query input to a classifier, our method prod
uces a progressive set of plausible variations of that query, which gradually ch
ange the posterior probability from its original class to its negation. These co
unter-factually generated samples preserve features unrelated to the classificat
ion decision, such that a user can employ our method as a ``tuning knob'' to tra
verse a data manifold while crossing the decision boundary. Our method is model
 agnostic and only requires the output value and gradient of the predictor with
respect to its input.
****

Quantum Optical Experiments Modeled by Long Short-Term Memory
Thomas Adler,Manuel Erhard,Mario Krenn,Johannes Brandstetter,Johannes Kofler,Sep
p Hochreiter
We demonstrate how machine learning is able to model experiments in quantum phys
ics. Quantum entanglement is a cornerstone for upcoming quantum technologies suc
h as quantum computation and quantum cryptography. Of particular interest are co
mplex quantum states with more than two particles and a large number of entangle
d quantum levels. Given such a multiparticle high-dimensional quantum state, it
is usually impossible to reconstruct an experimental setup that produces it. To
search for interesting experiments, one thus has to randomly create millions of

setups on a computer and calculate the respective output states. In this work, w e show that machine learning models can provide significant improvement over ran dom search. We demonstrate that a long short-term memory (LSTM) neural network c an successfully learn to model quantum experiments by correctly predicting outpu t state characteristics for given setups without the necessity of computing the states themselves. This approach not only allows for faster search but is also a n essential step towards automated design of multiparticle high-dimensional quan tum experiments using generative machine learning models.
****
Why do These Match? Explaining the Behavior of Image Similarity Models
Bryan A. Plummer,Mariya I. Vasileva,Vitali Petsiuk,Kate Saenko,David Forsyth
Explaining a deep learning model can help users understand its behavior and allo w researchers to discern its shortcomings. Recent work has primarily focused on explaining models for tasks like image classification or visual question answeri ng. In this paper, we introduce an explanation approach for image similarity mo dels, where a model's output is a score measuring the similarity of two inputs r ather than a classification. In this task, an explanation depends on both of th e input images, so standard methods do not apply. We propose an explanation meth od that pairs a saliency map identifying important image regions with an attribu te that best explains the match. We find that our explanations provide addition al information not typically captured by saliency maps alone, and can also impro ve performance on the classic task of attribute recognition. Our approach's abil ity to generalize is demonstrated on two datasets from diverse domains, Polyvore Outfits and Animals with Attributes 2.
****
Mode Connectivity and Sparse Neural Networks
Jonathan Frankle,Gintare Karolina Dziugaite,Daniel M. Roy,Michael Carbin
We uncover a connection between two seemingly unrelated empirical phenomena: mod e connectivity and sparsity. On the one hand, there is growing catalog of situat ions where, across multiple runs, SGD learns weights that fall into minima that are connected (mode connectivity). A striking example is described by Nagarajan & Kolter (2019). They observe that test error on MNIST does not change along the linear path connecting the end points of two independent SGD runs, starting fro m the same random initialization. On the other hand, there is the lottery ticket hypothesis of Frankle & Carbin (2019), where dense, randomly initialized networ ks have sparse subnetworks capable of training in isolation to full accuracy.

However, neither phenomenon scales beyond small vision networks. We start by pro posing a technique to find sparse subnetworks after initialization. We observe t hat these subnetworks match the accuracy of the full network only when two SGD r uns for the same subnetwork are connected by linear paths with the no change in test error. Our findings connect the existence of sparse subnetworks that train to high accuracy with the dynamics of optimization via mode connectivity. In doi ng so, we identify analogues of the phenomena uncovered by Nagarajan & Kolter an d Frankle & Carbin in ImageNet-scale architectures at state-of-the-art sparsity levels.
****
Monte Carlo Deep Neural Network Arithmetic
Julian Faraone,Philip Leong
Quantization is a crucial technique for achieving low-power, low latency and hig h throughput hardware implementations of Deep Neural Networks. Quantized floatin g point representations have received recent interest due to their hardware effi ciency benefits and ability to represent a higher dynamic range than fixed point representations, leading to improvements in accuracy. We present a novel techn ique, Monte Carlo Deep Neural Network Arithmetic (MCA), for determining the sens itivity of Deep Neural Networks to quantization in floating point arithmetic.We do this by applying Monte Carlo Arithmetic to the inference computation and anal yzing the relative standard deviation of the neural network loss. The method ma kes no assumptions regarding the underlying parameter distributions. We evaluate our method on pre-trained image classification models on the CIFAR10 andImageNe

t datasets. For the same network topology and dataset, we demonstrate the abili ty to gain the equivalent of bits of precision by simply choosing weight paramet er sets which demonstrate a lower loss of significance from the Monte Carlo tria ls. Additionally, we can apply MCA to compare the sensitivity of different ne twork topologies to quantization effects.
****

How can we generalise learning distributed representations of graphs?
Paul M Scherer,Pietro Lio
We propose a general framework to construct unsupervised models capable of learn ing distributed representations of discrete structures such as graphs based on R -Convolution kernels and distributed semantics research. Our framework combines the insights and observations of Deep Graph Kernels and Graph2Vec towards a unif ied methodology for performing similarity learning on graphs of arbitrary size. This is exemplified by our own instance G2DR which extends Graph2Vec from labell ed graphs towards unlabelled graphs and tackles issues of diagonal dominance thr ough pruning of the subgraph vocabulary composing graphs. These changes produce new state of the art results in the downstream application of G2DR embeddings in graph classification tasks over datasets with small labelled graphs in binary c lassification to multi-class classification on large unlabelled graphs using an off-the-shelf support vector machine.
****

Relation-based Generalized Zero-shot Classification with the Domain Discriminato r on the shared representation
Masahiro Suzuki,Yutaka Matsuo
Generalized zero-shot learning (GZSL) is the task of predicting a test image fro m seen or unseen classes using pre-defined class-attributes and images from the seen classes. Typical ZSL models assign the class corresponding to the most rel evant attribute as the predicted label of the test image based on the learned re lation between the attribute and the image. However, this relation-based approac h presents a difficulty: many of the test images are predicted as biased to the seen domain, i.e., the \emph{domain bias problem}. Recently, many methods have a ddressed this difficulty using a synthesis-based approach that, however, require s generation of large amounts of high-quality unseen images after training and t he additional training of classifier given them. Therefore, for this study, we aim at alleviating this difficulty in the manner of the relation-based approach. First, we consider the requirements for good performance in a ZSL setting and i ntroduce a new model based on a variational autoencoder that learns to embed att ributes and images into the shared representation space which satisfies those re quirements. Next, we assume that the domain bias problem in GZSL derives from a situation in which embedding of the unseen domain overlaps that of the seen one. We introduce a discriminator that distinguishes domains in a shared space and l earns jointly with the above embedding model to prevent this situation. After t raining, we can obtain prior knowledge from the discriminator of which domain is more likely to be embedded anywhere in the shared space. We propose combination of this knowledge and the relation-based classification on the embedded shared space as a mixture model to compensate class prediction. Experimentally obtained results confirm that the proposed method significantly improves the domain bias problem in relation-based settings and achieves almost equal accuracy to that o f high-cost synthesis-based methods.

****

Self-supervised Training of Proposal-based Segmentation via Background Predictio n
Isinsu Katircioglu,Helge Rhodin,Victor Constantin,Jörg Spörri,Mathieu Salzmann,P ascal Fua
While supervised object detection and segmentation methods achieve impressive ac curacy, they generalize poorly to images whose appearance significantly differs from the data they have been trained on. To address this in scenarios where anno tating data is prohibitively expensive, we introduce a self-supervised approach to detection and segmentation, able to work with monocular images captured with

a moving camera. At the heart of our approach lies the observations that object segmentation and background reconstruction are linked tasks, and that, for structured scenes, background regions can be re-synthesized from their surroundings, whereas regions depicting the object cannot.
We encode this intuition as a self-supervised loss function that we exploit to train a proposal-based segmentation network. To account for the discrete nature of the proposals, we develop a Monte Carlo-based training strategy that allows the algorithm to explore the large space of object proposals. We apply our method to human detection and segmentation in images that visually depart from those of standard benchmarks, achieving competitive results compared to the few existing self-supervised methods and approaching the accuracy of supervised ones that exploit large annotated datasets.
****

Decoupling Hierarchical Recurrent Neural Networks With Locally Computable Losses
Asier Mujika,Felix Weissenberger,Angelika Steger
Learning long-term dependencies is a key long-standing challenge of recurrent neural networks (RNNs). Hierarchical recurrent neural networks (HRNNs) have been considered a promising approach as long-term dependencies are resolved through shortcuts up and down the hierarchy. Yet, the memory requirements of Truncated Backpropagation Through Time (TBPTT) still prevent training them on very long sequences. In this paper, we empirically show that in (deep) HRNNs, propagating gradients back from higher to lower levels can be replaced by locally computable losses, without harming the learning capability of the network, over a wide range of tasks. This decoupling by local losses reduces the memory requirements of training by a factor exponential in the depth of the hierarchy in comparison to standard TBPTT.
****

Avoiding Negative Side-Effects and Promoting Safe Exploration with Imaginative Planning
Dhruv Ramani,Benjamin Eysenbach
With the recent proliferation of the usage of reinforcement learning (RL) agents for solving real-world tasks, safety emerges as a necessary ingredient for their successful application. In this paper, we focus on ensuring the safety of the agent while making sure that the agent does not cause any unnecessary disruptions to its environment. The current approaches to this problem, such as manually constraining the agent or adding a safety penalty to the reward function, can introduce bad incentives. In complex domains, these approaches are simply intractable, as they require knowing apriori all the possible unsafe scenarios an agent could encounter. We propose a model-based approach to safety that allows the agent to look into the future and be aware of the future consequences of its actions. We learn the transition dynamics of the environment and generate a directed graph called the imaginative module. This graph encapsulates all possible trajectories that can be followed by the agent, allowing the agent to efficiently traverse through the imagined environment without ever taking any action in reality. A baseline state, which can either represent a safe or an unsafe state (based on whichever is easier to define) is taken as a human input, and the imaginative module is used to predict whether the current actions of the agent can cause it to end up in dangerous states in the future. Our imaginative module can be seen as a ``plug-and-play'' approach to ensuring safety, as it is compatible with any existing RL algorithm and any task with discrete action space. Our method induces the agent to act safely while learning to solve the task. We experimentally validate our proposal on two gridworld environments and a self-driving car simulator, demonstrating that our approach to safety visits unsafe states significantly less frequently than a baseline.
****

BayesOpt Adversarial Attack
Binxin Ru,Adam Cobb,Arno Blaas,Yarin Gal
Black-box adversarial attacks require a large number of attempts before finding successful adversarial examples that are visually indistinguishable from the original input. Current approaches relying on substitute model training, gradient e

stimation or genetic algorithms often require an excessive number of queries. Th
erefore, they are not suitable for real-world systems where the maximum query nu
mber is limited due to cost. We propose a query-efficient black-box attack which
 uses Bayesian optimisation in combination with Bayesian model selection to opti
mise over the adversarial perturbation and the optimal degree of search space di
mension reduction. We demonstrate empirically that our method can achieve compar
able success rates with 2-5 times fewer queries compared to previous state-of-th
e-art black-box attacks.
****
CrossNorm: On Normalization for Off-Policy Reinforcement Learning
Aditya Bhatt,Max Argus,Artemij Amiranashvili,Thomas Brox
Off-policy temporal difference (TD) methods are a powerful class of reinforcemen
t learning (RL) algorithms. Intriguingly, deep off-policy TD algorithms are not
commonly used in combination with feature normalization techniques, despite posi
tive effects of normalization in other domains. We show that naive application o
f existing normalization techniques is indeed not effective, but that well-desig
ned normalization improves optimization stability and removes the necessity of t
arget networks. In particular, we introduce a normalization based on a mixture o
f on- and off-policy transitions, which we call cross-normalization. It can be r
egarded as an extension of batch normalization that re-centers data for two diff
erent distributions, as present in off-policy learning. Applied to DDPG and TD3,
 cross-normalization improves over the state of the art across a range of MuJoCo
 benchmark tasks.

****
A Simple Technique to Enable Saliency Methods to Pass the Sanity Checks
Arushi Gupta,Sanjeev Arora
 {\em Saliency methods} attempt to explain a deep net's decision by assigning a
{\em score} to each feature/pixel in the input, often doing this credit-assignme
nt via the gradient of the output with respect to input.
Recently \citet{adebayosan} questioned the validity of many of these methods sin
ce they do not pass simple {\em sanity checks}, which test whether the scores sh
ift/vanish when layers of the trained net are randomized, or when the net is re
trained using random labels for inputs. % for the inputs. %Surprisingly, the t
ested methods did not pass these checks: the explanations were relatively unchan
ged.

We propose a simple fix to existing saliency methods that helps them pass sanity
 checks, which we call {\em competition for pixels}. This involves computing sal
iency maps for all possible labels in the classification task, and using a simpl
e competition among them to identify and remove less relevant pixels from the ma
p. Some theoretical justification is provided for it and its performance is emp
irically demonstrated on several popular methods.
****
Directional Message Passing for Molecular Graphs
Johannes Gasteiger,Janek Groß,Stephan Günnemann
Graph neural networks have recently achieved great successes in predicting quant
um mechanical properties of molecules. These models represent a molecule as a gr
aph using only the distance between atoms (nodes). They do not, however, conside
r the spatial direction from one atom to another, despite directional informatio
n playing a central role in empirical potentials for molecules, e.g. in angular
potentials. To alleviate this limitation we propose directional message passing,
 in which we embed the messages passed between atoms instead of the atoms themse
lves. Each message is associated with a direction in coordinate space. These dir
ectional message embeddings are rotationally equivariant since the associated di
rections rotate with the molecule. We propose a message passing scheme analogous
 to belief propagation, which uses the directional information by transforming m
essages based on the angle between them. Additionally, we use spherical Bessel f
unctions and spherical harmonics to construct theoretically well-founded, orthog
onal representations that achieve better performance than the currently prevalen

t Gaussian radial basis representations while using fewer than 1/4 of the parame
ters. We leverage these innovations to construct the directional message passing
 neural network (DimeNet). DimeNet outperforms previous GNNs on average by 76% o
n MD17 and by 31% on QM9. Our implementation is available online.
****

Unsupervised Learning of Efficient and Robust Speech Representations
Kazuya Kawakami,Luyu Wang,Chris Dyer,Phil Blunsom,Aaron van den Oord
We present an unsupervised method for learning speech representations based on a
 bidirectional contrastive predictive coding that implicitly discovers phonetic
structure from large-scale corpora of unlabelled raw audio signals. The represen
tations, which we learn from up to 8000 hours of publicly accessible speech data
, are evaluated by looking at their impact on the behaviour of supervised speech
 recognition systems. First, across a variety of datasets, we find that the feat
ures learned from the largest and most diverse pretraining dataset result in sig
nificant improvements over standard audio features as well as over features lear
ned from smaller amounts of pretraining data. Second, they significantly improve
 sample efficiency in low-data scenarios. Finally, the features confer significa
nt robustness advantages to the resulting recognition systems: we see significan
t improvements in out-of-domain transfer relative to baseline feature sets, and
the features likewise provide improvements in four different low-resource Africa
n language datasets.
****

Compositional Embeddings: Joint Perception and Comparison of Class Label Sets
Zeqian Li,Jacob Whitehill
We explore the idea of compositional set embeddings that can be used to infer no
t
just a single class, but the set of classes associated with the input data (e.g.
, image,
video, audio signal). This can be useful, for example, in multi-object detection
 in
images, or multi-speaker diarization (one-shot learning) in audio. In particular
, we
devise and implement two novel models consisting of (1) an embedding function
f trained jointly with a “composite” function g that computes set union opera-
tions between the classes encoded in two embedding vectors; and (2) embedding
f trained jointly with a “query” function h that computes whether the classes en
-
coded in one embedding subsume the classes encoded in another embedding. In
contrast to prior work, these models must both perceive the classes associated
with the input examples, and also encode the relationships between different cla
ss
label sets. In experiments conducted on simulated data, OmniGlot, and COCO
datasets, the proposed composite embedding models outperform baselines based
on traditional embedding approaches.
****

Model-based reinforcement learning for biological sequence design
Christof Angermueller,David Dohan,David Belanger,Ramya Deshpande,Kevin Murphy,Lu
cy Colwell
The ability to design biological structures such as DNA or proteins would have c
onsiderable medical and industrial impact. Doing so presents a challenging black
-box optimization problem characterized by the large-batch, low round setting du
e to the need for labor-intensive wet lab evaluations. In response, we propose u
sing reinforcement learning (RL) based on proximal-policy optimization (PPO) for
 biological sequence design. RL provides a flexible framework for optimization g
enerative sequence models to achieve specific criteria, such as diversity among
the high-quality sequences discovered. We propose a model-based variant of PPO,
DyNA-PPO, to improve sample efficiency, where the policy for a new round is trai
ned offline using a simulator fit on functional measurements from prior rounds.
To accommodate the growing number of observations across rounds, the simulator m
odel is automatically selected at each round from a pool of diverse models of va

rying capacity. On the tasks of designing DNA transcription factor binding site
s, designing antimicrobial proteins, and optimizing the energy of Ising models b
ased on protein structure, we find that DyNA-PPO performs significantly better t
han existing methods in settings in which modeling is feasible, while still not
performing worse in situations in which a reliable model cannot be learned.
****
Learning to Optimize via Dual space Preconditioning
Sélim Chraibi,Adil Salim,Samuel Horváth,Filip Hanzely,Peter Richtárik
Preconditioning an minimization algorithm improve its convergence and can lead t
o a minimizer in one iteration in some extreme cases. There is currently no anal
ytical way for finding a suitable preconditioner. We present a general methodolo
gy for learning the preconditioner and show that it can lead to dramatic speed-u
ps over standard optimization techniques.

****
Self-Attentional Credit Assignment for Transfer in Reinforcement Learning
Johan Ferret,Raphaël Marinier,Matthieu Geist,Olivier Pietquin
The ability to transfer knowledge to novel environments and tasks is a sensible
desiderata for general learning agents. Despite the apparent promises, transfer
in RL is still an open and little exploited research area. In this paper, we tak
e a brand-new perspective about transfer: we suggest that the ability to assign
credit unveils structural invariants in the tasks that can be transferred to mak
e RL more sample efficient. Our main contribution is Secret, a novel approach to
 transfer learning for RL that uses a backward-view credit assignment mechanism
based on a self-attentive architecture. Two aspects are key to its generality: i
t learns to assign credit as a separate offline supervised process and exclusive
ly modifies the reward function. Consequently, it can be supplemented by transfe
r methods that do not modify the reward function and it can be plugged on top of
 any RL algorithm.
****
AdaGAN: Adaptive GAN for Many-to-Many Non-Parallel Voice Conversion
Maitreya Patel,Mirali Purohit,Mihir Parmar,Nirmesh J. Shah,Hemant A. Patil
Voice Conversion (VC) is a task of converting perceived speaker identity from a
source speaker to a particular target speaker. Earlier approaches in the literat
ure primarily find a mapping between the given source-target speaker-pairs. Deve
loping mapping techniques for many-to-many VC using non-parallel data, including
 zero-shot learning remains less explored areas in VC. Most of the many-to-many
VC architectures require training data from all the target speakers for whom we
want to convert the voices. In this paper, we propose a novel style transfer arc
hitecture, which can also be extended to generate voices even for target speaker
s whose data were not used in the training (i.e., case of zero-shot learning). I
n particular, propose Adaptive Generative Adversarial Network (AdaGAN), new arch
itectural training procedure help in learning normalized speaker-independent lat
ent representation, which will be used to generate speech with different speakin
g styles in the context of VC. We compare our results with the state-of-the-art
StarGAN-VC architecture. In particular, the AdaGAN achieves 31.73%, and 10.37% r
elative improvement compared to the StarGAN in MOS tests for speech quality and
speaker similarity, respectively. The key strength of the proposed architectures
 is that it yields these results with less computational complexity. AdaGAN is 8
8.6% less complex than StarGAN-VC in terms of FLoating Operation Per Second (FLO
PS), and 85.46% less complex in terms of trainable parameters.
****
City Metro Network Expansion with Reinforcement Learning
Yu Wei,Minjia Mao,Xi Zhao,Jianhua Zou
This paper presents a method to solve the city metro network expansion problem u
sing reinforcement learning (RL). In this method, we formulate the metro expansi
on as a process of sequential station selection, and design feasibility rules ba
sed on the selected station sequence to ensure the reasonable connection pattern
s of metro line. Following this formulation, we train an actor critic model to d
esign the next metro line. The actor is a seq2seq network with attention mechani

sm to generate the parameterized policy which is the probability distribution ov er feasible stations. The critic is used to estimate the expected reward, which is determined by the output station sequences generated by the actor during trai ning, in order to reduce the training variance. The learning procedure only requ ires the reward calculation, thus our general method can be extended to multi-fa ctor cases easily. Considering origin-destination (OD) trips and social equity, we expand the current metro network in Xi'an, China, based on the real mobility information of 24,770,715 mobile phone users in the whole city. The results demo nstrate the effectiveness of our method.
****

BinaryDuo: Reducing Gradient Mismatch in Binary Activation Network by Coupling B inary Activations
Hyungjun Kim,Kyungsu Kim,Jinseok Kim,Jae-Joon Kim
Binary Neural Networks (BNNs) have been garnering interest thanks to their compu te cost reduction and memory savings. However, BNNs suffer from performance degr adation mainly due to the gradient mismatch caused by binarizing activations. Pr evious works tried to address the gradient mismatch problem by reducing the disc repancy between activation functions used at forward pass and its differentiable approximation used at backward pass, which is an indirect measure. In this work , we use the gradient of smoothed loss function to better estimate the gradient mismatch in quantized neural network. Analysis using the gradient mismatch estim ator indicates that using higher precision for activation is more effective than modifying the differentiable approximation of activation function. Based on the observation, we propose a new training scheme for binary activation networks ca lled BinaryDuo in which two binary activations are coupled into a ternary activa tion during training. Experimental results show that BinaryDuo outperforms state -of-the-art BNNs on various benchmarks with the same amount of parameters and co mputing cost.
****

ShardNet: One Filter Set to Rule Them All
Saumya Jetley,Tommaso Cavallari,Philip Torr,Stuart Golodetz
Deep CNNs have achieved state-of-the-art performance for numerous machine learni ng and computer vision tasks in recent years, but as they have become increasing ly deep, the number of parameters they use has also increased, making them hard to deploy in memory-constrained environments and difficult to interpret. Machine learning theory implies that such networks are highly over-parameterised and th at it should be possible to reduce their size without sacrificing accuracy, and indeed many recent studies have begun to highlight specific redundancies that ca n be exploited to achieve this. In this paper, we take a further step in this di rection by proposing a filter-sharing approach to compressing deep CNNs that red uces their memory footprint by repeatedly applying a single convolutional mappin g of learned filters to simulate a CNN pipeline. We show, via experiments on CIF AR-10, CIFAR-100, Tiny ImageNet, and ImageNet that this allows us to reduce the parameter counts of networks based on common designs such as VGGNet and ResNet b y a factor proportional to their depth, whilst leaving their accuracy largely un affected. At a broader level, our approach also indicates how the scale-space re gularities found in visual signals can be leveraged to build neural architecture s that are more parsimonious and interpretable.
****

Towards Interpretable Evaluations: A Case Study of Named Entity Recognition
Jinlan Fu,Pengfei Liu,Xuanjing Huang
With the proliferation of models for natural language processing (NLP) tasks , it is even harder to understand the differences between models and their relat ive merits. Simply looking at differences between holistic metrics such as accur acy, BLEU, or F1 do not tell us \emph{why} or \emph{how} a particular method is better and how dataset biases influence the choices of model design.

In this paper, we present a general methodology for {\emph{interpretable}} e valuation of NLP systems and choose the task of named entity recognition (NER) a s a case study, which is a core task of identifying people, places, or organizat ions in text. The proposed evaluation method enables us to interpret the \textit

{model biases}, \textit{dataset biases}, and how the \emph{differences in the da
tasets} affect the design of the models, identifying the strengths and weaknesse
s of current approaches. By making our {analysis} tool available, we make it eas
y for future researchers to run similar analyses and drive the progress in this
area.
****

Mixed-curvature Variational Autoencoders
Ondrej Skopek,Octavian-Eugen Ganea,Gary Bécigneul
Euclidean space has historically been the typical workhorse geometry for machine
 learning applications due to its power and simplicity. However, it has recently
 been shown that geometric spaces with constant non-zero curvature improve repre
sentations and performance on a variety of data types and downstream tasks. Cons
equently, generative models like Variational Autoencoders (VAEs) have been succe
ssfully generalized to elliptical and hyperbolic latent spaces. While these appr
oaches work well on data with particular kinds of biases e.g. tree-like data for
 a hyperbolic VAE, there exists no generic approach unifying and leveraging all
three models. We develop a Mixed-curvature Variational Autoencoder, an efficient
 way to train a VAE whose latent space is a product of constant curvature Rieman
nian manifolds, where the per-component curvature is fixed or learnable. This ge
neralizes the Euclidean VAE to curved latent spaces and recovers it when curvatu
res of all latent space components go to 0.
****

Rethinking deep active learning: Using unlabeled data at model training
Oriane Siméoni,Mateusz Budnik,Yannis Avrithis,Guillaume Gravier
Active learning typically focuses on training a model on few labeled examples al
one, while unlabeled ones are only used for acquisition. In this work we depart
from this setting by using both labeled and unlabeled data during model training
 across active learning cycles. We do so by using unsupervised feature learning
at the beginning of the active learning pipeline and semi-supervised learning at
 every active learning cycle, on all available data. The former has not been inv
estigated before in active learning, while the study of latter in the context of
 deep learning is scarce and recent findings are not conclusive with respect to
its benefit. Our idea is orthogonal to acquisition strategies by using more data
, much like ensemble methods use more models. By systematically evaluating on a
number of popular acquisition strategies and datasets, we find that the use of u
nlabeled data during model training brings a spectacular accuracy improvement in
 image classification, compared to the differences between acquisition strategie
s. We thus explore smaller label budgets, even one label per class.
****

Blurring Structure and Learning to Optimize and Adapt Receptive Fields
Evan Shelhamer,Dequan Wang,Trevor Darrell
The visual world is vast and varied, but its variations divide into structured a
nd unstructured factors. We compose free-form filters and structured Gaussian fi
lters, optimized end-to-end, to factorize deep representations and learn both lo
cal features and their degree of locality. In effect this optimizes over recepti
ve field size and shape, tuning locality to the data and task. Our semi-structur
ed composition is strictly more expressive than free-form filtering, and changes
 in its structured parameters would require changes in architecture for standard
 networks. Dynamic inference, in which the Gaussian structure varies with the in
put, adapts receptive field size to compensate for local scale variation. Optimi
zing receptive field size improves semantic segmentation accuracy on Cityscapes
by 1-2 points for strong dilated and skip architectures and by up to 10 points f
or suboptimal designs. Adapting receptive fields by dynamic Gaussian structure f
urther improves results, equaling the accuracy of free-form deformation while im
proving efficiency.
****

Layerwise Learning Rates for Object Features in Unsupervised and Supervised Neur
al Networks And Consequent Predictions for the Infant Visual System
Rhodri Cusack,Cliona O'Doherty,Anna Birbeck,Anna Truzzi
To understand how object vision develops in infancy and childhood, it will be ne

cessary to develop testable computational models. Deep neural networks (DNNs) ha
ve proven valuable as models of adult vision, but it is not yet clear if they ha
ve any value as models of development. As a first model, we measured learning in
 a DNN designed to mimic the architecture and representational geometry of the v
isual system (CORnet). We quantified the development of explicit object represen
tations at each level of this network through training by freezing the convoluti
onal layers and training an additional linear decoding layer. We evaluate decodi
ng accuracy on the whole ImageNet validation set, and also for individual visual
 classes. CORnet, however, uses supervised training and because infants have onl
y extremely impoverished access to labels they must instead learn in an unsuperv
ised manner. We therefore also measured learning in a state-of-the-art unsupervi
sed network (DeepCluster). CORnet and DeepCluster differ in both supervision and
 in the convolutional networks at their heart, thus to isolate the effect of sup
ervision, we ran a control experiment in which we trained the convolutional netw
ork from DeepCluster (an AlexNet variant) in a supervised manner. We make predic
tions on how learning should develop across brain regions in infants. In all thr
ee networks, we also tested for a relationship in the order in which infants and
 machines acquire visual classes, and found only evidence for a counter-intuitiv
e relationship. We discuss the potential reasons for this.
****

Continual Deep Learning by Functional Regularisation of Memorable Past
Pingbo Pan,Alexander Immer,Siddharth Swaroop,Runa Eschenhagen,Richard E Turner,M
ohammad Emtiyaz Khan
Continually learning new skills without forgetting old ones is an important qual
ity for an intelligent system, yet most deep learning methods suffer from catast
rophic forgetting of the past. Recent works have addressed this by regularising
the network weights, but it is challenging to identify weights crucial to avoid
forgetting. A better approach is to directly regularise the network outputs at p
ast inputs, e.g., by using Gaussian processes (GPs), but this is usually computa
tionally challenging. In this paper, we propose a scalable functional-regularisa
tion approach where we regularise only over a few memorable past examples that a
re crucial to avoid forgetting. Our key idea is to use a GP formulation of deep
networks, enabling us to both identify the memorable past and regularise over th
em. Our method achieves state-of-the-art performance on standard benchmarks and
opens a new direction for life-long learning where regularisation methods are na
turally combined with memory-based methods.
****

Demystifying Inter-Class Disentanglement
Aviv Gabbay,Yedid Hoshen
Learning to disentangle the hidden factors of variations within a set of observa
tions is a key task for artificial intelligence. We present a unified formulatio
n for class and content disentanglement and use it to illustrate the limitations
 of current methods. We therefore introduce LORD, a novel method based on Latent
 Optimization for Representation Disentanglement. We find that latent optimizati
on, along with an asymmetric noise regularization, is superior to amortized infe
rence for achieving disentangled representations. In extensive experiments, our
method is shown to achieve better disentanglement performance than both adversar
ial and non-adversarial methods that use the same level of supervision. We furth
er introduce a clustering-based approach for extending our method for settings t
hat exhibit in-class variation with promising results on the task of domain tran
slation.
****

On the implicit minimization of alternative loss functions when training deep ne
tworks
Alexandre Lemire Paquin,Brahim Chaib-draa,Philippe Giguère
Understanding the implicit bias of optimization algorithms is important in order
 to improve generalization of neural networks. One approach to try to exploit su
ch understanding would be to then make the bias explicit in the loss function.
Conversely, an interesting approach to gain more insights into the implicit bias
 could be to study how different loss functions are being implicitly minimized

when training the network. In this work, we concentrate our study on the inductive bias occurring when minimizing the cross-entropy loss with different batch sizes and learning rates. We investigate how three loss functions are being implicitly minimized during training. These three loss functions are the Hinge loss with different margins, the cross-entropy loss with different temperatures and a newly introduced Gcdf loss with different standard deviations. This Gcdf loss establishes a connection between a sharpness measure for the 0–1 loss and margin based loss functions. We find that a common behavior is emerging for all the loss functions considered.
****

A Deep Recurrent Neural Network via Unfolding Reweighted l1-l1 Minimization
Huynh Van Luong,Duy Hung Le,Nikos Deligiannis
Deep unfolding methods design deep neural networks as learned variations of optimization methods. These networks have been shown to achieve faster convergence and higher accuracy than the original optimization methods. In this line of research, this paper develops a novel deep recurrent neural network (coined reweighted-RNN) by unfolding a reweighted l1-l1 minimization algorithm and applies it to the task of sequential signal reconstruction. To the best of our knowledge, this is the first deep unfolding method that explores reweighted minimization. Due to the underlying reweighted minimization model, our RNN has a different soft-thresholding function (alias, different activation function) for each hidden unit in each layer. Furthermore, it has higher network expressivity than existing deep unfolding RNN models due to the over-parameterizing weights. Moreover, we establish theoretical generalization error bounds for the proposed reweighted-RNN model by means of Rademacher complexity. The bounds reveal that the parameterization of the proposed reweighted-RNN ensures good generalization. We apply the proposed reweighted-RNN to the problem of video-frame reconstruction from low-dimensional measurements, that is, sequential frame reconstruction. The experimental results on the moving MNIST dataset demonstrate that the proposed deep reweighted-RNN significantly outperforms existing RNN models.
****

Differentially Private Mixed-Type Data Generation For Unsupervised Learning
Uthaipon Tantipongpipat,Chris Waites,Digvijay Boob,Amaresh Siva,Rachel Cummings
In this work we introduce the DP-auto-GAN framework for synthetic data generation, which combines the low dimensional representation of autoencoders with the flexibility of GANs. This framework can be used to take in raw sensitive data, and privately train a model for generating synthetic data that should satisfy the same statistical properties as the original data. This learned model can be used to generate arbitrary amounts of publicly available synthetic data, which can then be freely shared due to the post-processing guarantees of differential privacy. Our framework is applicable to unlabled \emph{mixed-type data}, that may include binary, categorical, and real-valued data. We implement this framework on both unlabeled binary data (MIMIC-III) and unlabeled mixed-type data (ADULT). We also introduce new metrics for evaluating the quality of synthetic mixed-type data, particularly in unsupervised settings.
****

Learning from Rules Generalizing Labeled Exemplars
Abhijeet Awasthi,Sabyasachi Ghosh,Rasna Goyal,Sunita Sarawagi
In many applications labeled data is not readily available, and needs to be collected via pain-staking human supervision. We propose a rule-exemplar method for collecting human supervision to combine the efficiency of rules with the quality of instance labels. The supervision is coupled such that it is both natural for humans and synergistic for learning. We propose a training algorithm that jointly denoises rules via latent coverage variables, and trains the model through a soft implication loss over the coverage and label variables. The denoised rules and trained model are used jointly for inference. Empirical evaluation on five different tasks shows that (1) our algorithm is more accurate than several existing methods of learning from a mix of clean and noisy supervision, and (2) the coupled rule-exemplar supervision is effective in denoising rules.
****

Group-Transformer: Towards A Lightweight Character-level Language Model
Sungrae Park,Geewook Kim,Junyeop Lee,Junbum Cha,Ji-Hoon Kim Hwalsuk Lee
Character-level language modeling is an essential but challenging task in Natura
l Language Processing.
Prior works have focused on identifying long-term dependencies between character
s and have built deeper and wider networks for better performance. However, thei
r models require substantial computational resources, which hinders the usabilit
y of character-level language models in applications with limited resources. In
this paper, we propose a lightweight model, called Group-Transformer, that reduc
es the resource requirements for a Transformer, a promising method for modeling
sequence with long-term dependencies. Specifically, the proposed method partitio
ns linear operations to reduce the number of parameters and computational cost.
As a result, Group-Transformer only uses 18.2\% of parameters compared to the be
st performing LSTM-based model, while providing better performance on two benchm
ark tasks, enwik8 and text8. When compared to Transformers with a comparable num
ber of parameters and time complexity, the proposed model shows better performan
ce. The implementation code will be available.
****
Language-independent Cross-lingual Contextual Representations
Xiao Zhang,Song Wang,Dejing Dou,Xien Liu,Thien Huu Nguyen,Ji Wu
Contextual representation models like BERT have achieved state-of-the-art perfor
mance on a diverse range of NLP tasks. We propose a cross-lingual contextual rep
resentation model that generates language-independent contextual representations
. This helps to enable zero-shot cross-lingual transfer of a wide range of NLP m
odels, on top of contextual representation models like BERT. We provide a formul
ation of language-independent cross-lingual contextual representation based on m
ono-lingual representations. Our formulation takes three steps to align sequence
s of vectors: transform, extract, and reorder. We present a detailed discussion
about the process of learning cross-lingual contextual representations, also abo
ut the performance in cross-lingual transfer learning and its implications.
****
Understanding the Limitations of Conditional Generative Models
Ethan Fetaya,Joern-Henrik Jacobsen,Will Grathwohl,Richard Zemel
Class-conditional generative models hold promise to overcome the shortcomings of
 their discriminative counterparts. They are a natural choice to solve discrimin
ative tasks in a robust manner as they jointly optimize for predictive performan
ce and accurate modeling of the input distribution. In this work, we investigate
 robust classification with likelihood-based generative models from a theoretica
l and practical perspective to investigate if they can deliver on their promises
. Our analysis focuses on a spectrum of robustness properties: (1) Detection of
worst-case outliers in the form of adversarial examples; (2) Detection of averag
e-case outliers in the form of ambiguous inputs and (3) Detection of incorrectly
 labeled in-distribution inputs.

Our theoretical result reveals that it is impossible to guarantee detectability
of adversarially-perturbed inputs even for near-optimal generative classifiers.
Experimentally, we find that while we are able to train robust models for MNIST,
 robustness completely breaks down on CIFAR10. We relate this failure to various
 undesirable model properties that can be traced to the maximum likelihood train
ing objective. Despite being a common choice in the literature, our results indi
cate that likelihood-based conditional generative models may are surprisingly in
effective for robust classification.
****
Skew-Explore: Learn faster in continuous spaces with sparse rewards
Xi Chen,Yuan Gao,Ali Ghadirzadeh,Marten Bjorkman,Ginevra Castellano,Patric Jensf
elt
In many reinforcement learning settings, rewards which are extrinsically availab
le to the learning agent are too sparse to train a suitable policy. Beside rewar
d shaping which requires human expertise, utilizing better exploration strategie
s helps to circumvent the problem of policy training with sparse rewards. In thi

s work, we introduce an exploration approach based on maximizing the entropy of the visited states while learning a goal-conditioned policy. The main contributi on of this work is to introduce a novel reward function which combined with a go al proposing scheme, increases the entropy of the visited states faster compared to the prior work. This improves the exploration capability of the agent, and t herefore enhances the agent's chance to solve sparse reward problems more effici ently. Our empirical studies demonstrate the superiority of the proposed method to solve different sparse reward problems in comparison to the prior work.
****
Exploring the Correlation between Likelihood of Flow-based Generative Models and Image Semantics
Xin WANG,SiuMing Yiu
Among deep generative models, flow-based models, simply referred as \emph{flow} s in this paper, differ from other models in that they provide tractable likelih ood. Besides being an evaluation metric of synthesized data, flows are supposed to be robust against out-of-distribution~(OoD) inputs since they do not discard any information of the inputs. However, it has been observed that flows trained on FashionMNIST assign higher likelihoods to OoD samples from MNIST. This counte r-intuitive observation raises the concern about the robustness of flows' likeli hood. In this paper, we explore the correlation between flows' likelihood and im age semantics. We choose two typical flows as the target models: Glow, based on coupling transformations, and pixelCNN, based on autoregressive transformations. Our experiments reveal surprisingly weak correlation between flows' likelihoods and image semantics: the predictive likelihoods of flows can be heavily affecte d by trivial transformations that keep the image semantics unchanged, which we c all semantic-invariant transformations~(SITs). We explore three SITs~(all small pixel-level modifications): image pixel translation, random noise perturbation, latent factors zeroing~(limited to flows using multi-scale architecture, e.g. Gl ow). These findings, though counter-intuitive, resonate with the fact that the p redictive likelihood of a flow is the joint probability of all the image pixels. So flows' likelihoods, modeling on pixel-level intensities, is not able to indi cate the existence likelihood of the high-level image semantics. We call for att ention that it may be \emph{abuse} if we use the predictive likelihoods of flows for OoD samples detection.
****
Anomaly Detection Based on Unsupervised Disentangled Representation Learning in Combination with Manifold Learning
Xiaoyan Li,Iluju Kiringa,Tet Yeap,Xiaodan Zhu,Yifeng Li
Identifying anomalous samples from highly complex and unstructured data is a cru cial but challenging task in a variety of intelligent systems. In this paper, we present a novel deep anomaly detection framework named AnoDM (standing for Anom aly detection based on unsupervised Disentangled representation learning and Man ifold learning). The disentanglement learning is currently implemented by beta-V AE for automatically discovering interpretable factorized latent representations in a completely unsupervised manner. The manifold learning is realized by t-SNE for projecting the latent representations to a 2D map. We define a new anomaly score function by combining beta-VAE's reconstruction error in the raw feature space and local density estimation in the t-SNE space. AnoDM was evaluated on bo th image and time-series data and achieved better results than models that use j ust one of the two measures and other deep learning methods.
****
Neural Arithmetic Unit by reusing many small pre-trained networks
Ammar Ahmad,Oneeb Babar,Murtaza Taj
We propose a solution for evaluation of mathematical expression. However, instea d of designing a single end-to-end model we propose a Lego bricks style architec ture. In this architecture instead of training a complex end-to-end neural netwo rk, many small networks can be trained independently each accomplishing one spec ific operation and acting a single lego brick. More difficult or complex task ca n then be solved using a combination of these smaller network. In this work we f irst identify 8 fundamental operations that are commonly used to solve arithmeti

c operations (such as 1 digit multiplication, addition, subtraction, sign calcul ator etc). These fundamental operations are then learned using simple feed forwa rd neural networks. We then shows that different operations can be designed simp ly by reusing these smaller networks. As an example we reuse these smaller netwo rks to develop larger and a more complex network to solve n-digit multiplication , n-digit division, and cross product. This bottom-up strategy not only introduc es reusability, we also show that it allows to generalize for computations invol ving n-digits and we show results for up to 7 digit numbers. Unlike existing met hods, our solution also generalizes for both positive as well as negative number s.
****

On Stochastic Sign Descent Methods
Mher Safaryan,Peter Richtárik
Various gradient compression schemes have been proposed to mitigate the communic ation cost in distributed training of large scale machine learning models. Sign-based methods, such as signSGD (Bernstein et al., 2018), have recently been gain ing popularity because of their simple compression rule and connection to adapti ve gradient methods, like ADAM. In this paper, we perform a general analysis of sign-based methods for non-convex optimization. Our analysis is built on intuiti ve bounds on success probabilities and does not rely on special noise distributi ons nor on the boundedness of the variance of stochastic gradients. Extending th e theory to distributed setting within a parameter server framework, we assure e xponentially fast variance reduction with respect to number of nodes, maintainin g 1-bit compression in both directions and using small mini-batch sizes. We vali date our theoretical findings experimentally.
****

GENN: Predicting Correlated Drug-drug Interactions with Graph Energy Neural Netw orks
Tengfei Ma,Junyuan Shang,Cao Xiao,Jimeng Sun
Gaining more comprehensive knowledge about drug-drug interactions (DDIs) is one of the most important tasks in drug development and medical practice. Recently g raph neural networks have achieved great success in this task by modeling drugs as nodes and drug-drug interactions as links and casting DDI predictions as link prediction problems. However, correlations between link labels (e.g., DDI types) were rarely considered in existing works.

We propose the graph energy neural network (\mname) to explicitly model link ty pe correlations. We formulate the DDI prediction task as a structure prediction problem and introduce a new energy-based model where the energy function is def ined by graph neural networks. Experiments on two real-world DDI datasets demons trated that \mname is superior to many baselines without consideration of link t ype correlations and achieved 13.77% and 5.01% PR-AUC improvement on the t wo datasets, respectively. We also present a case study in which \mname can bett er capture meaningful DDI correlations compared with baseline models.
****

Event Discovery for History Representation in Reinforcement Learning
Aleksandr Ermolov,Enver Sangineto,Nicu Sebe
Environments in Reinforcement Learning (RL) are usually only partially observabl e. To address this problem, a possible solution is to provide the agent with inf ormation about past observations. While common methods represent this history u sing a Recurrent Neural Network (RNN), in this paper we propose an alternative r epresentation which is based on the record of the past events observed in a give n episode. Inspired by the human memory, these events describe only important ch anges in the environment and, in our approach, are automatically discovered usin g self-supervision.

We evaluate our history representation method using two challenging RL benchmar ks: some games of the Atari-57 suite and the 3D environment Obstacle Tower. Usin g these benchmarks we show the advantage of our solution with respect to common RNN-based approaches.
****

Keep Doing What Worked: Behavior Modelling Priors for Offline Reinforcement Lear

ning
Noah Siegel,Jost Tobias Springenberg,Felix Berkenkamp,Abbas Abdolmaleki,Michael Neunert,Thomas Lampe,Roland Hafner,Nicolas Heess,Martin Riedmiller
Off-policy reinforcement learning algorithms promise to be applicable in settings where only a fixed data-set (batch) of environment interactions is available and no new experience can be acquired. This property makes these algorithms appealing for real world problems such as robot control. In practice, however, standard off-policy algorithms fail in the batch setting for continuous control. In this paper, we propose a simple solution to this problem. It admits the use of data generated by arbitrary behavior policies and uses a learned prior -- the advantage-weighted behavior model (ABM) -- to bias the RL policy towards actions that have previously been executed and are likely to be successful on the new task. Our method can be seen as an extension of recent work on batch-RL that enables stable learning from conflicting data-sources. We find improvements on competitive baselines in a variety of RL tasks -- including standard continuous control benchmarks and multi-task learning for simulated and real-world robots.
****
Are Powerful Graph Neural Nets Necessary? A Dissection on Graph Classification
Ting Chen,Song Bian,Yizhou Sun
Graph Neural Nets (GNNs) have received increasing attentions, partially due to their superior performance in many node and graph classification tasks. However, there is a lack of understanding on what they are learning and how sophisticated the learned graph functions are. In this work, we propose a dissection of GNNs on graph classification into two parts: 1) the graph filtering, where graph-based neighbor aggregations are performed, and 2) the set function, where a set of hidden node features are composed for prediction. To study the importance of both parts, we propose to linearize them separately. We first linearize the graph filtering function, resulting Graph Feature Network (GFN), which is a simple lightweight neural net defined on a \textit{set} of graph augmented features. Further linearization of GFN's set function results in Graph Linear Network (GLN), which is a linear function. Empirically we perform evaluations on common graph classification benchmarks. To our surprise, we find that, despite the simplification, GFN could match or exceed the best accuracies produced by recently proposed GNNs (with a fraction of computation cost), while GLN underperforms significantly. Our results demonstrate the importance of non-linear set function, and suggest that linear graph filtering with non-linear set function is an efficient and powerful scheme for modeling existing graph classification benchmarks.
****
Domain-Invariant Representations: A Look on Compression and Weights
Victor Bouvier,Céline Hudelot,Clément Chastagnol,Philippe Very,Myriam Tami
Learning Invariant Representations to adapt deep classifiers of a source domain to a new target domain has recently attracted much attention. In this paper, we show that the search for invariance favors the compression of representations. We point out this may have a bad impact on adaptability of representations expressed as a minimal combined domain error. By considering the risk of compression, we show that weighting representations can align representation distributions without impacting their adaptability. This supports the claim that representation invariance is too strict a constraint. First, we introduce a new bound on the target risk that reveals a trade-off between compression and invariance of learned representations. More precisely, our results show that the adaptability of a representation can be better controlled when the compression risk is taken into account. In contrast, preserving adaptability may overestimate the risk of compression that makes the bound impracticable. We support these statements with a theoretical analysis illustrated on a standard domain adaptation benchmark. Second, we show that learning weighted representations plays a key role in relaxing the constraint of invariance and then preserving the risk of compression. Taking advantage of this trade-off may open up promising directions for the design of new adaptation methods.
****
Minimally distorted Adversarial Examples with a Fast Adaptive Boundary Attack

Francesco Croce,Matthias Hein
The evaluation of robustness against adversarial manipulations of neural networks-based classifiers is mainly tested with empirical attacks as the methods for the exact computation, even when available, do not scale to large networks. We propose in this paper a new white-box adversarial attack wrt the l_p-norms for $p \in \{1,2,\infty\}$ aiming at finding the minimal perturbation necessary to change the class of a given input. It has an intuitive geometric meaning, yields quickly high quality results, minimizes the size of the perturbation (so that it returns the robust accuracy at every threshold with a single run). It performs better or similarly to state-of-the-art attacks which are partially specialized to one l_p-norm.
****
Empirical Bayes Transductive Meta-Learning with Synthetic Gradients
Shell Xu Hu,Pablo Garcia Moreno,Yang Xiao,Xi Shen,Guillaume Obozinski,Neil Lawrence,Andreas Damianou
We propose a meta-learning approach that learns from multiple tasks in a transductive setting, by leveraging the unlabeled query set in addition to the support set to generate a more powerful model for each task. To develop our framework, we revisit the empirical Bayes formulation for multi-task learning. The evidence lower bound of the marginal log-likelihood of empirical Bayes decomposes as a sum of local KL divergences between the variational posterior and the true posterior on the query set of each task.
We derive a novel amortized variational inference that couples all the variational posteriors via a meta-model, which consists of a synthetic gradient network and an initialization network. Each variational posterior is derived from synthetic gradient descent to approximate the true posterior on the query set, although where we do not have access to the true gradient.
Our results on the Mini-ImageNet and CIFAR-FS benchmarks for episodic few-shot classification outperform previous state-of-the-art methods. Besides, we conduct two zero-shot learning experiments to further explore the potential of the synthetic gradient.
****
Spike-based causal inference for weight alignment
Jordan Guerguiev,Konrad Kording,Blake Richards
In artificial neural networks trained with gradient descent, the weights used for processing stimuli are also used during backward passes to calculate gradients. For the real brain to approximate gradients, gradient information would have to be propagated separately, such that one set of synaptic weights is used for processing and another set is used for backward passes. This produces the so-called "weight transport problem" for biological models of learning, where the backward weights used to calculate gradients need to mirror the forward weights used to process stimuli. This weight transport problem has been considered so hard that popular proposals for biological learning assume that the backward weights are simply random, as in the feedback alignment algorithm. However, such random weights do not appear to work well for large networks. Here we show how the discontinuity introduced in a spiking system can lead to a solution to this problem. The resulting algorithm is a special case of an estimator used for causal inference in econometrics, regression discontinuity design. We show empirically that this algorithm rapidly makes the backward weights approximate the forward weights. As the backward weights become correct, this improves learning performance over feedback alignment on tasks such as Fashion-MNIST and CIFAR-10. Our results demonstrate that a simple learning rule in a spiking network can allow neurons to produce the right backward connections and thus solve the weight transport problem.
****
Symmetry and Systematicity
Jeff Mitchell,Jeff Bowers
We argue that symmetry is an important consideration in addressing the problem of systematicity and investigate two forms of symmetry relevant to symbolic processes.

We implement this approach in terms of convolution and show that it can
be used to achieve effective generalisation in three toy problems: rule learning
,
composition and grammar learning.
****
Efficacy of Pixel-Level OOD Detection for Semantic Segmentation
Matt Angus,Krzysztof Czarnecki,Rick Salay
The detection of out of distribution samples for image classification has been w
idely researched. Safety critical applications, such as autonomous driving, woul
d benefit from the ability to localise the unusual objects causing the image to
be out of distribution. This paper adapts state-of-the-art methods for detecting
 out of distribution images for image classification to the new task of detectin
g out of distribution pixels, which can localise the unusual objects. It further
 experimentally compares the adapted methods on two new datasets derived from ex
isting semantic segmentation datasets using PSPNet and DeeplabV3+ architectures,
 as well as proposing a new metric for the task. The evaluation shows that the p
erformance ranking of the compared methods does not transfer to the new task and
 every method performs significantly worse than their image-level counterparts.
****
PatchFormer: A neural architecture for self-supervised representation learning o
n images
Aravind Srinivas,Pieter Abbeel
Learning rich representations from predictive learning without labels has been a
 longstanding challenge in the field of machine learning. Generative pre-trainin
g has so far not been as successful as contrastive methods in modeling represent
ations of raw images. In this paper, we propose a neural architecture for self-s
upervised representation learning on raw images called the PatchFormer which lea
rns to model spatial dependencies across patches in a raw image. Our method lear
ns to model the conditional probability distribution of missing patches given th
e context of surrounding patches. We evaluate the utility of the learned represe
ntations by fine-tuning the pre-trained model on low data-regime classification
tasks. Specifically, we benchmark our model on semi-supervised ImageNet classifi
cation which has become a popular benchmark recently for semi-supervised and sel
f-supervised learning methods. Our model is able to achieve 30.3% and 65.5% top-
1 accuracies when trained only using 1% and 10% of the labels on ImageNet showin
g the promise for generative pre-training methods.
****
Address2vec: Generating vector embeddings for blockchain analytics
Ali Hussein,Samiiha Nalwooga
Bitcoin is a virtual coinage system that enables users to trade virtually free o
f a central trusted authority. All transactions on the Bitcoin blockchain are pu
blicly available for viewing, yet as Bitcoin is built mainly for security it’s o
riginal structure does not allow for direct analysis of address transactions.
Existing analysis methods of the Bitcoin blockchain can be complicated, computat
ionally expensive or inaccurate. We propose a computationally efficient model to
 analyze bitcoin blockchain addresses and allow for their use with existing mach
ine learning algorithms. We compare our approach against Multi Level Sequence Le
arners (MLSLs), one of the best performing models on bitcoin address data.
****
Attack-Resistant Federated Learning with Residual-based Reweighting
Shuhao Fu,Chulin Xie,Bo Li,Qifeng Chen
Federated learning has a variety of applications in multiple domains by utilizin
g private training data stored on different devices. However, the aggregation pr
ocess in federated learning is highly vulnerable to adversarial attacks so that
the global model may behave abnormally under attacks. To tackle this challenge,
we present a novel aggregation algorithm with residual-based reweighting to defe
nd federated learning. Our aggregation algorithm combines repeated median regres
sion with the reweighting scheme in iteratively reweighted least squares. Our ex
periments show that our aggression algorithm outperforms other alternative algor
ithms in the presence of label-flipping, backdoor, and Gaussian noise attacks. W

e also provide theoretical guarantees for our aggregation algorithm.

****
Learning scalable and transferable multi-robot/machine sequential assignment pla nning via graph embedding
Hyunwook Kang,Aydar Mynbay,James R. Morrison,Jinkyoo Park
Can the success of reinforcement learning methods for simple combinatorial optim ization problems be extended to multi-robot sequential assignment planning? In a ddition to the challenge of achieving near-optimal performance in large problems , transferability to an unseen number of robots and tasks is another key challen ge for real-world applications. In this paper, we suggest a method that achieves the first success in both challenges for robot/machine scheduling problems.

Our method comprises of three components. First, we show any robot scheduling pr oblem can be expressed as a random probabilistic graphical model (PGM). We devel op a mean-field inference method for random PGM and use it for Q-function infere nce. Second, we show that transferability can be achieved by carefully designing two-step sequential encoding of problem state. Third, we resolve the computatio nal scalability issue of fitted Q-iteration by suggesting a heuristic auction-ba sed Q-iteration fitting method enabled by transferability we achieved.

We apply our method to discrete-time, discrete space problems (Multi-Robot Rewar d Collection (MRRC)) and scalably achieve 97% optimality with transferability. T his optimality is maintained under stochastic contexts. By extending our method to continuous time, continuous space formulation, we claim to be the first learn ing-based method with scalable performance in any type of multi-machine scheduli ng problems; our method scalability achieves comparable performance to popular m etaheuristics in Identical parallel machine scheduling (IPMS) problems.
****
Learning a Spatio-Temporal Embedding for Video Instance Segmentation
Anthony Hu,Alex Kendall,Roberto Cipolla
Understanding object motion is one of the core problems in computer vision. It r equires segmenting and tracking objects over time. Significant progress has been made in instance segmentation, but such models cannot track objects, and more c rucially, they are unable to reason in both 3D space and time.
We propose a new spatio-temporal embedding loss on videos that generates tempora lly consistent video instance segmentation. Our model includes a temporal networ k that learns to model temporal context and motion, which is essential to produc e smooth embeddings over time. Further, our model also estimates monocular depth , with a self-supervised loss, as the relative distance to an object effectively constrains where it can be next, ensuring a time-consistent embedding. Finally, we show that our model can accurately track and segment instances, even with oc clusions and missed detections, advancing the state-of-the-art on the KITTI Mult i-Object and Tracking Dataset.
****
Efficient Exploration via State Marginal Matching
Lisa Lee,Benjain Eysenbach,Emilio Parisotto,Erix Xing,Sergey Levine,Ruslan Salak hutdinov
Reinforcement learning agents need to explore their unknown environments to solv e the tasks given to them. The Bayes optimal solution to exploration is intracta ble for complex environments, and while several exploration methods have been pr oposed as approximations, it remains unclear what underlying objective is being optimized by existing exploration methods, or how they can be altered to incorpo rate prior knowledge about the task. Moreover, it is unclear how to acquire a si ngle exploration strategy that will be useful for solving multiple downstream ta sks. We address these shortcomings by learning a single exploration policy that can quickly solve a suite of downstream tasks in a multi-task setting, amortizin g the cost of learning to explore. We recast exploration as a problem of State M arginal Matching (SMM), where we aim to learn a policy for which the state margi nal distribution matches a given target state distribution, which can incorporat

e prior knowledge about the task. We optimize the objective by reducing it to a two-player, zero-sum game between a state density model and a parametric policy. Our theoretical analysis of this approach suggests that prior exploration methods do not learn a policy that does distribution matching, but acquire a replay buffer that performs distribution matching, an observation that potentially explains these prior methods' success in single-task settings. On both simulated and real-world tasks, we demonstrate that our algorithm explores faster and adapts more quickly than prior methods.
****
Lookahead: A Far-sighted Alternative of Magnitude-based Pruning
Sejun Park*,Jaeho Lee*,Sangwoo Mo,Jinwoo Shin
Magnitude-based pruning is one of the simplest methods for pruning neural networks. Despite its simplicity, magnitude-based pruning and its variants demonstrated remarkable performances for pruning modern architectures. Based on the observation that magnitude-based pruning indeed minimizes the Frobenius distortion of a linear operator corresponding to a single layer, we develop a simple pruning method, coined lookahead pruning, by extending the single layer optimization to a multi-layer optimization. Our experimental results demonstrate that the proposed method consistently outperforms magnitude-based pruning on various networks, including VGG and ResNet, particularly in the high-sparsity regime. See https://github.com/alinlab/lookahead_pruning for codes.
****
SCELMo: Source Code Embeddings from Language Models
Rafael - Michael Karampatsis,Charles Sutton
Continuous embeddings of tokens in computer programs have been used to support a variety of software development tools, including readability, code search, and program repair.
Contextual embeddings are common in natural language processing but have not been previously applied in software engineering.
We introduce a new set of deep contextualized word representations for computer programs based on language models.
We train a set of embeddings using the ELMo (embeddings from language models) framework of Peters et al (2018).
We investigate whether these embeddings are effective when fine-tuned for the downstream task of bug detection.
We show that even a low-dimensional embedding trained on a relatively small corpus of programs can improve a state-of-the-art machine learning system for bug detection.
****
Detecting Change in Seasonal Pattern via Autoencoder and Temporal Regularization
Raphael Fettaya,Dor Bank,Rachel Lemberg,Linoy Barel
Change-point detection problem consists of discovering abrupt property changes in the generation process of time-series. Most state-of-the-art models are optimizing the power of a kernel two-sample test, with only a few assumptions on the distribution of the data. Unfortunately, because they presume the samples are distributed i.i.d, they are not able to use information about the seasonality of a time-series. In this paper, we present a novel approach - ATR-CSPD allowing the detection of changes in the seasonal pattern of a time-series. Our method uses an autoencoder together with a temporal regularization, to learn the pattern of each seasonal cycle. Using low dimensional representation of the seasonal patterns, it is possible to accurately and efficiently estimate the existence of a change point using a clustering algorithm. Through experiments on artificial and real-world data sets, we demonstrate the usefulness of the proposed method for several applications.
****
VariBAD: A Very Good Method for Bayes-Adaptive Deep RL via Meta-Learning
Luisa Zintgraf,Kyriacos Shiarlis,Maximilian Igl,Sebastian Schulze,Yarin Gal,Katja Hofmann,Shimon Whiteson
Trading off exploration and exploitation in an unknown environment is key to maximising expected return during learning. A Bayes-optimal policy, which does so o

ptimally, conditions its actions not only on the environment state but on the ag ent’s uncertainty about the environment. Computing a Bayes-optimal policy is how ever intractable for all but the smallest tasks. In this paper, we introduce var iational Bayes-Adaptive Deep RL (variBAD), a way to meta-learn to perform approx imate inference in an unknown environment, and incorporate task uncer- tainty di rectly during action selection. In a grid-world domain, we illustrate how variBA D performs structured online exploration as a function of task uncertainty. We f urther evaluate variBAD on MuJoCo domains widely used in meta-RL and show that i t achieves higher online return than existing methods.
****

A Generalized Training Approach for Multiagent Learning
Paul Muller,Shayegan Omidshafiei,Mark Rowland,Karl Tuyls,Julien Perolat,Siqi Liu ,Daniel Hennes,Luke Marris,Marc Lanctot,Edward Hughes,Zhe Wang,Guy Lever,Nicolas Heess,Thore Graepel,Remi Munos
This paper investigates a population-based training regime based on game-theoret ic principles called Policy-Spaced Response Oracles (PSRO). PSRO is general in t he sense that it (1) encompasses well-known algorithms such as fictitious play a nd double oracle as special cases, and (2) in principle applies to general-sum, many-player games. Despite this, prior studies of PSRO have been focused on two- player zero-sum games, a regime where in Nash equilibria are tractably computabl e. In moving from two-player zero-sum games to more general settings, computatio n of Nash equilibria quickly becomes infeasible. Here, we extend the theoretica l underpinnings of PSRO by considering an alternative solution concept, α-Rank, which is unique (thus faces no equilibrium selection issues, unlike Nash) and ap plies readily to general-sum, many-player settings. We establish convergence gua rantees in several games classes, and identify links between Nash equilibria and α-Rank. We demonstrate the competitive performance of α-Rank-based PSRO against an exact Nash solver-based PSRO in 2-player Kuhn and Leduc Poker. We then go be yond the reach of prior PSRO applications by considering 3- to 5-player poker ga mes, yielding instances where α-Rank achieves faster convergence than approximat e Nash solvers, thus establishing it as a favorable general games solver. We als o carry out an initial empirical validation in MuJoCo soccer, illustrating the f easibility of the proposed approach in another complex domain.
****

Quantum Semi-Supervised Kernel Learning
Seyran Saeedi,Aliakbar Panahi,Tom Arodz
Quantum machine learning methods have the potential to facilitate learning using extremely large datasets. While the availability of data for training machine l earning models is steadily increasing, oftentimes it is much easier to collect f eature vectors that to obtain the corresponding labels. One of the approaches fo r addressing this issue is to use semi-supervised learning, which leverages not only the labeled samples, but also unlabeled feature vectors. Here, we present a quantum machine learning algorithm for training Semi-Supervised Kernel Support Vector Machines. The algorithm uses recent advances in quantum sample-based Hami ltonian simulation to extend the existing Quantum LS-SVM algorithm to handle the semi-supervised term in the loss, while maintaining the same quantum speedup as the Quantum LS-SVM.
****

Unsupervised Meta-Learning for Reinforcement Learning
Abhishek Gupta,Benjamin Eysenbach,Chelsea Finn,Sergey Levine
Meta-learning algorithms learn to acquire new tasks more quickly from past exper ience. In the context of reinforcement learning, meta-learning algorithms can ac quire reinforcement learning procedures to solve new problems more efficiently b y utilizing experience from prior tasks. The performance of meta-learning algori thms depends on the tasks available for meta-training: in the same way that supe rvised learning generalizes best to test points drawn from the same distribution as the training points, meta-learning methods generalize best to tasks from the same distribution as the meta-training tasks. In effect, meta-reinforcement lea rning offloads the design burden from algorithm design to task design. If we can automate the process of task design as well, we can devise a meta-learning algo

rithm that is truly automated. In this work, we take a step in this direction, p
roposing a family of unsupervised meta-learning algorithms for reinforcement lea
rning. We motivate and describe a general recipe for unsupervised meta-reinforce
ment learning, and present an instantiation of this approach. Our conceptual and
 theoretical contributions consist of formulating the unsupervised meta-reinforc
ement learning problem and describing how task proposals based on mutual informa
tion can in principle be used to train optimal meta-learners. Our experimental r
esults indicate that unsupervised meta-reinforcement learning effectively acquir
es accelerated reinforcement learning procedures without the need for manual tas
k design and significantly exceeds the performance of learning from scratch.
****

Making Efficient Use of Demonstrations to Solve Hard Exploration Problems
Caglar Gulcehre,Tom Le Paine,Bobak Shahriari,Misha Denil,Matt Hoffman,Hubert Soy
er,Richard Tanburn,Steven Kapturowski,Neil Rabinowitz,Duncan Williams,Gabriel Ba
rth-Maron,Ziyu Wang,Nando de Freitas,Worlds Team
This paper introduces R2D3, an agent that makes efficient use of demonstrations
to solve hard exploration problems in partially observable environments with hig
hly variable initial conditions. We also introduce a suite of eight tasks that c
ombine these three properties, and show that R2D3 can solve several of the tasks
 where other state of the art methods (both with and without demonstrations) fai
l to see even a single successful trajectory after tens of billions of steps of
exploration.
****

Training individually fair ML models with sensitive subspace robustness
Mikhail Yurochkin,Amanda Bower,Yuekai Sun
We consider training machine learning models that are fair in the sense that the
ir performance is invariant under certain sensitive perturbations to the inputs.
 For example, the performance of a resume screening system should be invariant u
nder changes to the gender and/or ethnicity of the applicant. We formalize this
notion of algorithmic fairness as a variant of individual fairness and develop a
 distributionally robust optimization approach to enforce it during training. We
 also demonstrate the effectiveness of the approach on two ML tasks that are sus
ceptible to gender and racial biases.
****

Meta-learning curiosity algorithms
Ferran Alet*,Martin F. Schneider*,Tomas Lozano-Perez,Leslie Pack Kaelbling
We hypothesize that curiosity is a mechanism found by evolution that encourages
meaningful exploration early in an agent's life in order to expose it to experie
nces that enable it to obtain high rewards over the course of its lifetime. We f
ormulate the problem of generating curious behavior as one of meta-learning: an
outer loop will search over a space of curiosity mechanisms that dynamically ada
pt the agent's reward signal, and an inner loop will perform standard reinforcem
ent learning using the adapted reward signal. However, current meta-RL methods b
ased on transferring neural network weights have only generalized between very s
imilar tasks. To broaden the generalization, we instead propose to meta-learn al
gorithms: pieces of code similar to those designed by humans in ML papers. Our r
ich language of programs combines neural networks with other building blocks suc
h as buffers, nearest-neighbor modules and custom loss functions. We demonstrate
 the effectiveness of the approach empirically, finding two novel curiosity algo
rithms that perform on par or better than human-designed published curiosity alg
orithms in domains as disparate as grid navigation with image inputs, acrobot, l
unar lander, ant and hopper.
****

vq-wav2vec: Self-Supervised Learning of Discrete Speech Representations
Alexei Baevski,Steffen Schneider,Michael Auli
We propose vq-wav2vec to learn discrete representations of audio segments throug
h a wav2vec-style self-supervised context prediction task. The algorithm uses ei
ther a gumbel softmax or online k-means clustering to quantize the dense represe
ntations. Discretization enables the direct application of algorithms from the N
LP community which require discrete inputs. Experiments show that BERT pre-train

ing achieves a new state of the art on TIMIT phoneme classification and WSJ spee
ch recognition.
****
Leveraging Entanglement Entropy for Deep Understanding of Attention Matrix in T
ext Matching
Peng Zhang,XiaoLiu Mao,XinDian Ma,BenYou Wang,Jing Zhang,Jun Wang,DaWei Song
The formal understanding of deep learning has made great progress based on quant
um many-body physics. For example, the entanglement entropy in quantum many-bod
y systems can interpret the inductive bias of neural network and then guide the
design of network structure and parameters for certain tasks. However, there are
 two unsolved problems in the current study of entanglement entropy, which limit
s its application potential. First, the theoretical benefits of entanglement ent
ropy was only investigated in the representation of a single object (e.g., an im
age or a sentence), but has not been well studied in the matching of two objects
 (e.g., question-answering pairs). Second, the entanglement entropy can not be
qualitatively calculated since the exponentially increasing dimension of the ma
tching matrix. In this paper, we are trying to address these two problem by inve
stigating the fundamental connections between the entanglement entropy and the a
ttention matrix. We prove that by a mapping (via the trace operator) on the high
-dimensional matching matrix, a low-dimensional attention matrix can be derived
. Based on such a attention matrix, we can provide a feasible solution to the en
tanglement entropy that describes the correlation between the two objects in mat
ching tasks. Inspired by the theoretical property of the entanglement entropy, w
e can design the network architecture adaptively in a typical text matching task
, i.e., question-answering task.
****
Infinite-horizon Off-Policy Policy Evaluation with Multiple Behavior Policies
Xinyun Chen,Lu Wang,Yizhe Hang,Heng Ge,Hongyuan Zha
We consider off-policy policy evaluation when the trajectory data are generated
by multiple behavior policies. Recent work has shown the key role played by the
state or state-action stationary distribution corrections in the infinite horizo
n context for off-policy policy evaluation. We propose estimated mixture policy
(EMP), a novel class of partially policy-agnostic methods to accurately estimate
 those quantities. With careful analysis, we show that EMP gives rise to estimat
es with reduced variance for estimating the state stationary distribution correc
tion while it also offers a useful induction bias for estimating the state-actio
n stationary distribution correction. In extensive experiments with both continu
ous and discrete environments, we demonstrate that our algorithm offers signific
antly improved accuracy compared to the state-of-the-art methods.
****
Under what circumstances do local codes emerge in feed-forward neural networks
Ella M. Gale,Nicolas Martin
Localist coding schemes are more easily interpretable than the distributed schem
es but generally believed to be biologically implausible. Recent results have fo
und highly selective units and object detectors in NNs that are indicative of lo
cal codes (LCs). Here we undertake a constructionist study on feed-forward NNs a
nd find LCs emerging in response to invariant features, and this finding is robu
st until the invariant feature is perturbed by 40%. Decreasing the number of inp
ut data, increasing the relative weight of the invariant features and large valu
es of dropout all increase the number of LCs. Longer training times increase the
 number of LCs and the turning point of the LC-epoch curve correlates well with
the point at which NNs reach 90-100% on both test and training accuracy. Pseudo-
deep networks (2 hidden layers) which have many LCs lose them when common aspect
s of deep-NN research are applied (large training data, ReLU activations, early
stopping on training accuracy and softmax), suggesting that LCs may not be found
 in deep-NNs. Switching to more biologically feasible constraints (sigmoidal act
ivation functions, longer training times, dropout, activation noise) increases t
he number of LCs. If LCs are not found in the feed-forward classification layers
 of modern deep-CNNs these data suggest this could either be caused by a lack of
 (moderately) invariant features being passed to the fully connected layers or d

ue to the choice of training conditions and architecture. Should the interpretability and resilience to noise of LCs be required, this work suggests how to tune a NN so they emerge.
****
MMA Training: Direct Input Space Margin Maximization through Adversarial Training
Gavin Weiguang Ding,Yash Sharma,Kry Yik Chau Lui,Ruitong Huang
We study adversarial robustness of neural networks from a margin maximization perspective, where margins are defined as the distances from inputs to a classifier's decision boundary.
Our study shows that maximizing margins can be achieved by minimizing the adversarial loss on the decision boundary at the "shortest successful perturbation", demonstrating a close connection between adversarial losses and the margins. We propose Max-Margin Adversarial (MMA) training to directly maximize the margins to achieve adversarial robustness.
Instead of adversarial training with a fixed ϵ, MMA offers an improvement by enabling adaptive selection of the "correct" ϵ as the margin individually for each datapoint. In addition, we rigorously analyze adversarial training with the perspective of margin maximization, and provide an alternative interpretation for adversarial training, maximizing either a lower or an upper bound of the margins. Our experiments empirically confirm our theory and demonstrate MMA training's efficacy on the MNIST and CIFAR10 datasets w.r.t. ℓ_∞ and ℓ_2 robustness.
****
Forecasting Deep Learning Dynamics with Applications to Hyperparameter Tuning
Piotr Kozakowski,■ukasz Kaiser,Afroz Mohiuddin
Well-performing deep learning models have enormous impact, but getting them
to perform well is complicated, as the model architecture must be chosen and a
number of hyperparameters tuned. This requires experimentation, which is timeconsuming and costly. We propose to address the problem of hyperparameter
tuning by learning to forecast the training behaviour of deep learning architectures.
Concretely, we introduce a forecasting model that, given a hyperparameter schedule
(e.g., learning rate, weight decay) and a history of training observations (such as
loss and accuracy), predicts how the training will continue. Naturally, forecasting
is much faster and less expensive than running actual deep learning experiments.
The main question we study is whether the forecasting model is good enough to be
of use - can it indeed replace real experiments? We answer this affirmatively in two
ways. For one, we show that the forecasted curves are close to real ones. On the
practical side, we apply our forecaster to learn hyperparameter tuning policies. We
experiment on a version of ResNet on CIFAR10 and on Transformer in a language
modeling task. The policies learned using our forecaster match or exceed the ones
learned in real experiments and in one case even the default schedules discovered
by researchers. We study the learning rate schedules created using the forecaster
are find that they are not only effective, but also lead to interesting insights.
****
Batch Normalization has Multiple Benefits: An Empirical Study on Residual Networks
Soham De,Samuel L Smith
Many state of the art models rely on two architectural innovations; skip connections and batch normalization. However batch normalization has a number of limita

tions. It breaks the independence between training examples within a batch, perf orms poorly when the batch size is too small, and significantly increases the co st of computing a parameter update in some models. This work identifies two prac tical benefits of batch normalization. First, it improves the final test accurac y. Second, it enables efficient training with larger batches and larger learning rates. However we demonstrate that the increase in the largest stable learning rate does not explain why the final test accuracy is increased under a finite ep och budget. Furthermore, we show that the gap in test accuracy between residual networks with and without batch normalization can be dramatically reduced by imp roving the initialization scheme. We introduce “ZeroInit”, which trains a 1000 l ayer deep Wide-ResNet without normalization to 94.3% test accuracy on CIFAR-10 i n 200 epochs at batch size 64. This initialization scheme outperforms batch norm alization when the batch size is very small, and is competitive with batch norma lization for batch sizes that are not too large. We also show that ZeroInit matc hes the validation accuracy of batch normalization when training ResNet-50-V2 on ImageNet at batch size 1024.

****

Building Deep Equivariant Capsule Networks

Sai Raam Venkataraman,S. Balasubramanian,R. Raghunatha Sarma

Capsule networks are constrained by the parameter-expensive nature of their laye rs, and the general lack of provable equivariance guarantees. We present a varia tion of capsule networks that aims to remedy this. We identify that learning all pair-wise part-whole relationships between capsules of successive layers is ine fficient. Further, we also realise that the choice of prediction networks and th e routing mechanism are both key to equivariance. Based on these, we propose an alternative framework for capsule networks that learns to projectively encode th e manifold of pose-variations, termed the space-of-variation (SOV), for every ca psule-type of each layer. This is done using a trainable, equivariant function d efined over a grid of group-transformations. Thus, the prediction-phase of routi ng involves projection into the SOV of a deeper capsule using the corresponding function. As a specific instantiation of this idea, and also in order to reap th e benefits of increased parameter-sharing, we use type-homogeneous group-equivar iant convolutions of shallower capsules in this phase. We also introduce an equi variant routing mechanism based on degree-centrality. We show that this particul ar instance of our general model is equivariant, and hence preserves the composi tional representation of an input under transformations. We conduct several expe riments on standard object-classification datasets that showcase the increased t ransformation-robustness, as well as general performance, of our model to severa l capsule baselines.

****

Learning to Infer User Interface Attributes from Images

Philippe Schlattner,Pavol Bielik,Martin Vechev

We present a new approach that helps developers automate the process of user int erface implementation. Concretely, given an input image created by a designer (e .g, using a vector graphics editor), we learn to infer its implementation which when rendered (e.g., on the Android platform), looks visually the same as the in put image. To achieve this, we take a black box rendering engine and a set of at tributes it supports (e.g., colors, border radius, shadow or text properties), u se it to generate a suitable synthetic training dataset, and then train speciali zed neural models to predict each of the attribute values. To improve pixel-leve l accuracy, we also use imitation learning to train a neural policy that refines the predicted attribute values by learning to compute the similarity of the ori ginal and rendered images in their attribute space, rather than based on the dif ference of pixel values.

****

Attacking Graph Convolutional Networks via Rewiring

Yao Ma,Suhang Wang,Tyler Derr,Lingfei Wu,Jiliang Tang

Graph Neural Networks (GNNs) have boosted the performance of many graph related tasks such as node classification and graph classification. Recent researches sh

ow that graph neural networks are vulnerable to adversarial attacks, which delib
erately add carefully created unnoticeable perturbation to the graph structure.
The perturbation is usually created by adding/deleting a few edges, which might
be noticeable even when the number of edges modified is small. In this paper, we
 propose a graph rewiring operation which affects the graph in a less noticeable
 way compared to adding/deleting edges. We then use reinforcement learning to le
arn the attack strategy based on the proposed rewiring operation. Experiments on
 real world graphs demonstrate the effectiveness of the proposed framework. To u
nderstand the proposed framework, we further analyze how its generated perturbat
ion to the graph structure affects the output of the target model.
****

Incorporating BERT into Neural Machine Translation
Jinhua Zhu,Yingce Xia,Lijun Wu,Di He,Tao Qin,Wengang Zhou,Houqiang Li,Tieyan Liu
The recently proposed BERT (Devlin et al., 2019) has shown great power on a vari
ety of natural language understanding tasks, such as text classification, readin
g comprehension, etc. However, how to effectively apply BERT to neural machine t
ranslation (NMT) lacks enough exploration. While BERT is more commonly used as f
ine-tuning instead of contextual embedding for downstream language understanding
 tasks, in NMT, our preliminary exploration of using BERT as contextual embeddin
g is better than using for fine-tuning. This motivates us to think how to better
 leverage BERT for NMT along this direction. We propose a new algorithm named BE
RT-fused model, in which we first use BERT to extract representations for an inp
ut sequence, and then the representations are fused with each layer of the encod
er and decoder of the NMT model through attention mechanisms. We conduct experim
ents on supervised (including sentence-level and document-level translations), s
emi-supervised and unsupervised machine translation, and achieve state-of-the-ar
t results on seven benchmark datasets. Our code is available at https://github.c
om/bert-nmt/bert-nmt
****

Unsupervised Hierarchical Graph Representation Learning with Variational Bayes
Shashanka Ubaru,Jie Chen
Hierarchical graph representation learning is an emerging subject owing to the i
ncreasingly popular adoption of graph neural networks in machine learning and ap
plications. Loosely speaking, work under this umbrella falls into two categories
: (a) use a predefined graph hierarchy to perform pooling; and (b) learn the hie
rarchy for a given graph through differentiable parameterization of the coarseni
ng process. These approaches are supervised; a predictive task with ground-truth
 labels is used to drive the learning. In this work, we propose an unsupervised
approach, \textsc{BayesPool}, with the use of variational Bayes. It produces gra
ph representations given a predefined hierarchy. Rather than relying on labels,
the training signal comes from the evidence lower bound of encoding a graph and
decoding the subsequent one in the hierarchy. Node features are treated latent i
n this variational machinery, so that they are produced as a byproduct and are u
sed in downstream tasks. We demonstrate a comprehensive set of experiments to sh
ow the usefulness of the learned representation in the context of graph classifi
cation.
****

Copy That! Editing Sequences by Copying Spans
Sheena Panthaplackel,Miltiadis Allamanis,Marc Brockschmidt
Neural sequence-to-sequence models are finding increasing use in editing of docu
ments, for example in correcting a text document or repairing source code. In th
is paper, we argue that existing seq2seq models (with a facility to copy single
tokens) are not a natural fit for such tasks, as they have to explicitly copy ea
ch unchanged token. We present an extension of seq2seq models capable of copying
 entire spans of the input to the output in one step, greatly reducing the numbe
r of decisions required during inference. This extension means that there are no
w many ways of generating the same output, which we handle by deriving a new obj
ective for training and a variation of beam search for inference that explicitly
 handle this problem.

In our experiments on a range of editing tasks of natural language and source co
de, we show that our new model consistently outperforms simpler baselines.
****
DeepXML: Scalable & Accurate Deep Extreme Classification for Matching User Queri
es to Advertiser Bid Phrases
Kunal Dahiya,Anshul Mittal,Deepak Saini,Kushal Dave,Himanshu Jain,Sumeet Agarwal
,Manik Varma
The objective in deep extreme multi-label learning is to jointly learn feature r
epresentations and classifiers to automatically tag data points with the most re
levant subset of labels from an extremely large label set. Unfortunately, state-
of-the-art deep extreme classifiers are either not scalable or inaccurate for sh
ort text documents. This paper develops the DeepXML algorithm which addresses b
oth limitations by introducing a novel architecture that splits training of head
 and tail labels. DeepXML increases accuracy by (a) learning word embeddings on
 head labels and transferring them through a novel residual connection to data i
mpoverished tail labels; (b) increasing the amount of negative training data ava
ilable by extending state-of-the-art negative sub-sampling techniques; and (c) r
e-ranking the set of predicted labels to eliminate the hardest negatives for the
 original classifier. All of these contributions are implemented efficiently by
extending the highly scalable Slice algorithm for pretrained embeddings to learn
 the proposed DeepXML architecture. As a result, DeepXML could efficiently scale
 to problems involving millions of labels that were beyond the pale of state-of-
the-art deep extreme classifiers as it could be more than 10x faster at training
 than XML-CNN and AttentionXML. At the same time, DeepXML was also empirically d
etermined to be up to 19% more accurate than leading techniques for matching sea
rch engine queries to advertiser bid phrases.
****
What Can Neural Networks Reason About?
Keyulu Xu,Jingling Li,Mozhi Zhang,Simon S. Du,Ken-ichi Kawarabayashi,Stefanie Je
gelka
Neural networks have succeeded in many reasoning tasks. Empirically, these tasks
 require specialized network structures, e.g., Graph Neural Networks (GNNs) perf
orm well on many such tasks, but less structured networks fail. Theoretically, t
here is limited understanding of why and when a network structure generalizes be
tter than others, although they have equal expressive power. In this paper, we d
evelop a framework to characterize which reasoning tasks a network can learn wel
l, by studying how well its computation structure aligns with the algorithmic st
ructure of the relevant reasoning process. We formally define this algorithmic a
lignment and derive a sample complexity bound that decreases with better alignme
nt. This framework offers an explanation for the empirical success of popular re
asoning models, and suggests their limitations. As an example, we unify seemingl
y different reasoning tasks, such as intuitive physics, visual question answerin
g, and shortest paths, via the lens of a powerful algorithmic paradigm, dynamic
programming (DP). We show that GNNs align with DP and thus are expected to solve
 these tasks. On several reasoning tasks, our theory is supported by empirical r
esults.
****
Structured Object-Aware Physics Prediction for Video Modeling and Planning
Jannik Kossen,Karl Stelzner,Marcel Hussing,Claas Voelcker,Kristian Kersting
When humans observe a physical system, they can easily locate components, unders
tand their interactions, and anticipate future behavior, even in settings with c
omplicated and previously unseen interactions. For computers, however, learning
such models from videos in an unsupervised fashion is an unsolved research probl
em. In this paper, we present STOVE, a novel state-space model for videos, whi
ch explicitly reasons about objects and their positions, velocities, and interac
tions. It is constructed by combining an image model and a dynamics model in com
positional manner and improves on previous work by reusing the dynamics model fo
r inference, accelerating and regularizing training. STOVE predicts videos with
convincing physical behavior over hundreds of timesteps, outperforms previous un
supervised models, and even approaches the performance of supervised baselines.

We further demonstrate the strength of our model as a simulator for sample effic
ient model-based control, in a task with heavily interacting objects.

****
A multi-task U-net for segmentation with lazy labels
Rihuan Ke,Aurélie Bugeau,Nicolas Papadakis,Peter Schuetz,Carola-Bibiane Schönlie
b
The need for labour intensive pixel-wise annotation is a major limitation of man
y fully supervised learning methods for image segmentation. In this paper, we
 propose a deep convolutional neural network for multi-class segmentation that c
ircumvents this problem by being trainable on coarse data labels combined with o
nly a very small number of images with pixel-wise annotations. We call this new
 labelling strategy ‘lazy’ labels. Image segmentation is then stratified into t
hree connected tasks: rough detection of class instances, separation of wrongly
connected objects without a clear boundary, and pixel-wise segmentation to find
the accurate boundaries of each object. These problems are integrated into a mul
ti-task learning framework and the model is trained end-to-end in a semi-supervi
sed fashion. The method is demonstrated on two segmentation datasets, including
food microscopy images and histology images of tissues respectively. We show tha
t the model gives accurate segmentation results even if exact boundary labels ar
e missing for a majority of the annotated data. This allows more flexibility an
d efficiency for training deep neural networks that are data hungry in a practic
al setting where manual annotation is expensive, by collecting more lazy (rough)
 annotations than precisely segmented images.
****
Neural Design of Contests and All-Pay Auctions using Multi-Agent Simulation
Thomas Anthony,Ian Gemp,Janos Kramar,Tom Eccles,Andrea Tacchetti,Yoram Bachrach
We propose a multi-agent learning approach for designing crowdsourcing contests
and all-pay auctions. Prizes in contests incentivise contestants to expend effor
t on their entries, with different prize allocations resulting in different ince
ntives and bidding behaviors. In contrast to auctions designed manually by econo
mists, our method searches the possible design space using a simulation of the m
ulti-agent learning process, and can thus handle settings where a game-theoretic
 equilibrium analysis is not tractable. Our method simulates agent learning in c
ontests and evaluates the utility of the resulting outcome for the auctioneer. G
iven a large contest design space, we assess through simulation many possible co
ntest designs within the space, and fit a neural network to predict outcomes for
 previously untested contest designs. Finally, we apply mirror descent to optimi
ze the design so as to achieve more desirable outcomes. Our empirical analysis s
hows our approach closely matches the optimal outcomes in settings where the equ
ilibrium is known, and can produce high quality designs in settings where the eq
uilibrium strategies are not solvable analytically.
****
CaptainGAN: Navigate Through Embedding Space For Better Text Generation
Chun-Hsing Lin,Alvin Chiang,Chi-Liang Liu,Chien-Fu Lin,Po-Hsien Chu,Siang-Ruei W
u,Yi-En Tsai,Chung-Yang (Ric) Huang
Score-function-based text generation approaches such as REINFORCE, in general, s
uffer from high computational complexity and training instability problems. This
 is mainly due to the non-differentiable nature of the discrete space sampling a
nd thus these methods have to treat the discriminator as a reward function and i
gnore the gradient information. In this paper, we propose a novel approach, Capt
ainGAN, which adopts the straight-through gradient estimator and introduces a ”r
e-centered” gradient estimation technique to steer the generator toward better t
ext tokens through the embedding space. Our method is stable to train and conver
ges quickly without maximum likelihood pre-training. On multiple metrics of text
 quality and diversity, our method outperforms existing GAN-based methods on nat
ural language generation.
****
Learning-Augmented Data Stream Algorithms
Tanqiu Jiang,Yi Li,Honghao Lin,Yisong Ruan,David P. Woodruff

The data stream model is a fundamental model for processing massive data sets wi th limited memory and fast processing time. Recently Hsu et al. (2019) incorpora ted machine learning techniques into the data stream model in order to learn rel evant patterns in the input data. Such techniques were encapsulated by training an oracle to predict item frequencies in the streaming model. In this paper we e xplore the full power of such an oracle, showing that it can be applied to a wid e array of problems in data streams, sometimes resulting in the first optimal bo unds for such problems. Namely, we apply the oracle to counting distinct element s on the difference of streams, estimating frequency moments, estimating cascade d aggregates, and estimating moments of geometric data streams. For the distinct elements problem, we obtain the first memory-optimal algorithms. For estimating the p-th frequency moment for $0 < p < 2$ we obtain the first algorithms with optimal update time. For estimating the p-the frequency moment for $p > 2$ we obtain a quadratic saving in memory. We empirically validate our results, demon strating also our improvements in practice.

****

word2ket: Space-efficient Word Embeddings inspired by Quantum Entanglement

Aliakbar Panahi,Seyran Saeedi,Tom Arodz

Deep learning natural language processing models often use vector word embedding s, such as word2vec or GloVe, to represent words. A discrete sequence of words c an be much more easily integrated with downstream neural layers if it is represe nted as a sequence of continuous vectors. Also, semantic relationships between words, learned from a text corpus, can be encoded in the relative configurations of the embedding vectors. However, storing and accessing embedding vectors for all words in a dictionary requires large amount of space, and may stain systems with limited GPU memory. Here, we used approaches inspired by quantum computing to propose two related methods, word2ket and word2ketXS, for storing word embedd ing matrix during training and inference in a highly efficient way. Our approach achieves a hundred-fold or more reduction in the space required to store the em beddings with almost no relative drop in accuracy in practical natural language processing tasks.

****

On Weight-Sharing and Bilevel Optimization in Architecture Search

Mikhail Khodak,Liam Li,Maria-Florina Balcan,Ameet Talwalkar

Weight-sharing—the simultaneous optimization of multiple neural networks using t he same parameters—has emerged as a key component of state-of-the-art neural arc hitecture search. However, its success is poorly understood and often found to b e surprising. We argue that, rather than just being an optimization trick, the w eight-sharing approach is induced by the relaxation of a structured hypothesis s pace, and introduces new algorithmic and theoretical challenges as well as appli cations beyond neural architecture search. Algorithmically, we show how the geom etry of ERM for weight-sharing requires greater care when designing gradient- ba sed minimization methods and apply tools from non-convex non-Euclidean optimizat ion to give general-purpose algorithms that adapt to the underlying structure. W e further analyze the learning-theoretic behavior of the bilevel optimization so lved by practical weight-sharing methods. Next, using kernel configuration and N LP feature selection as case studies, we demonstrate how weight-sharing applies to the architecture search generalization of NAS and effectively optimizes the r esulting bilevel objective. Finally, we use our optimization analysis to develop a simple exponentiated gradient method for NAS that aligns with the underlying optimization geometry and matches state-of-the-art approaches on CIFAR-10.

****

Towards Hierarchical Importance Attribution: Explaining Compositional Semantics for Neural Sequence Models

Xisen Jin,Zhongyu Wei,Junyi Du,Xiangyang Xue,Xiang Ren

The impressive performance of neural networks on natural language processing tas ks attributes to their ability to model complicated word and phrase compositions . To explain how the model handles semantic compositions, we study hierarchical explanation of neural network predictions. We identify non-additivity and contex t independent importance attributions within hierarchies as two desirable proper

ties for highlighting word and phrase compositions. We show some prior efforts o n hierarchical explanations, e.g. contextual decomposition, do not satisfy the d esired properties mathematically, leading to inconsistent explanation quality in different models. In this paper, we start by proposing a formal and general way to quantify the importance of each word and phrase. Following the formulation, we propose Sampling and Contextual Decomposition (SCD) algorithm and Sampling an d Occlusion (SOC) algorithm. Human and metrics evaluation on both LSTM models an d BERT Transformer models on multiple datasets show that our algorithms outperfo rm prior hierarchical explanation algorithms. Our algorithms help to visualize s emantic composition captured by models, extract classification rules and improve human trust of models.

****

Compositional Transfer in Hierarchical Reinforcement Learning

Markus Wulfmeier,Abbas Abdolmaleki,Roland Hafner,Jost Tobias Springenberg,Michae l Neunert,Tim Hertweck,Thomas Lampe,Noah Siegel,Nicolas Heess,Martin Riedmiller

The successful application of flexible, general learning algorithms to real-worl d robotics applications is often limited by their poor data-efficiency. To addre ss the challenge, domains with more than one dominant task of interest encourage the sharing of information across tasks to limit required experiment time. To t his end, we investigate compositional inductive biases in the form of hierarchic al policies as a mechanism for knowledge transfer across tasks in reinforcement learning (RL). We demonstrate that this type of hierarchy enables positive trans fer while mitigating negative interference. Furthermore, we demonstrate the bene fits of additional incentives to efficiently decompose task solutions. Our exper iments show that these incentives are naturally given in multitask learning and can be easily introduced for single objectives. We design an RL algorithm that e nables stable and fast learning of structured policies and the effective reuse o f both behavior components and transition data across tasks in an off-policy set ting. Finally, we evaluate our algorithm in simulated environments as well as ph ysical robot experiments and demonstrate substantial improvements in data data-efficiency over competitive baselines.

****

On the Relationship between Self-Attention and Convolutional Layers

Jean-Baptiste Cordonnier,Andreas Loukas,Martin Jaggi

Recent trends of incorporating attention mechanisms in vision have led researche rs to reconsider the supremacy of convolutional layers as a primary building blo ck. Beyond helping CNNs to handle long-range dependencies, Ramachandran et al. (2019) showed that attention can completely replace convolution and achieve state -of-the-art performance on vision tasks. This raises the question: do learned at tention layers operate similarly to convolutional layers? This work provides evi dence that attention layers can perform convolution and, indeed, they often lear n to do so in practice. Specifically, we prove that a multi-head self-attention layer with sufficient number of heads is at least as expressive as any convoluti onal layer. Our numerical experiments then show that self-attention layers atten d to pixel-grid patterns similarly to CNN layers, corroborating our analysis. Ou r code is publicly available.

****

Dynamic Scale Inference by Entropy Minimization

Dequan Wang,Evan Shelhamer,Bruno Olshausen,Trevor Darrell

Given the variety of the visual world there is not one true scale for recognitio n: objects may appear at drastically different sizes across the visual field. Ra ther than enumerate variations across filter channels or pyramid levels, dynamic models locally predict scale and adapt receptive fields accordingly. The degree of variation and diversity of inputs makes this a difficult task. Existing meth ods either learn a feedforward predictor, which is not itself totally immune to the scale variation it is meant to counter, or select scales by a fixed algorith m, which cannot learn from the given task and data. We extend dynamic scale infe rence from feedforward prediction to iterative optimization for further adaptivi ty. We propose a novel entropy minimization objective for inference and optimize over task and structure parameters to tune the model to each input. Optimizatio

n during inference improves semantic segmentation accuracy and generalizes bette
r to extreme scale variations that cause feedforward dynamic inference to falter
.
****

SpikeGrad: An ANN-equivalent Computation Model for Implementing Backpropagation
with Spikes
Johannes C. Thiele,Olivier Bichler,Antoine Dupret
Event-based neuromorphic systems promise to reduce the energy consumption of dee
p neural networks by replacing expensive floating point operations on dense matr
ices by low energy, sparse operations on spike events. While these systems can b
e trained increasingly well using approximations of the backpropagation algorith
m, this usually requires high precision errors and is therefore incompatible wit
h the typical communication infrastructure of neuromorphic circuits. In this wor
k, we analyze how the gradient can be discretized into spike events when trainin
g a spiking neural network. To accelerate our simulation, we show that using a s
pecial implementation of the integrate-and-fire neuron allows us to describe the
 accumulated activations and errors of the spiking neural network in terms of an
 equivalent artificial neural network, allowing us to largely speed up training
compared to an explicit simulation of all spike events. This way we are able to
demonstrate that even for deep networks, the gradients can be discretized suffic
iently well with spikes if the gradient is properly rescaled. This form of spike
-based backpropagation enables us to achieve equivalent or better accuracies on
the MNIST and CIFAR10 datasets than comparable state-of-the-art spiking neural n
etworks trained with full precision gradients. The algorithm, which we call Spik
eGrad, is based on only accumulation and comparison operations and can naturally
 exploit sparsity in the gradient computation, which makes it an interesting cho
ice for a spiking neuromorphic systems with on-chip learning capacities.
****

GENERALIZATION GUARANTEES FOR NEURAL NETS VIA HARNESSING THE LOW-RANKNESS OF JAC
OBIAN
Samet Oymak,Zalan Fabian,Mingchen Li,Mahdi Soltanolkotabi
Modern neural network architectures often generalize well despite containing man
y more parameters than the size of the training dataset. This paper explores the
 generalization capabilities of neural networks trained via gradient descent. We
 develop a data-dependent optimization and generalization theory which leverages
 the low-rank structure of the Jacobian matrix associated with the network. Our
results help demystify why training and generalization is easier on clean and st
ructured datasets and harder on noisy and unstructured datasets as well as how t
he network size affects the evolution of the train and test errors during traini
ng. Specifically, we use a control knob to split the Jacobian spectum into ``inf
ormation" and ``nuisance" spaces associated with the large and small singular va
lues. We show that over the information space learning is fast and one can quick
ly train a model with zero training loss that can also generalize well. Over the
 nuisance space training is slower and early stopping can help with generalizati
on at the expense of some bias. We also show that the overall generalization cap
ability of the network is controlled by how well the labels are aligned with the
 information space. A key feature of our results is that even constant width neu
ral nets can provably generalize for sufficiently nice datasets. We conduct vari
ous numerical experiments on deep networks that corroborate our theoretical find
ings and demonstrate that: (i) the Jacobian of typical neural networks exhibit l
ow-rank structure with a few large singular values and many small ones leading t
o a low-dimensional information space, (ii) over the information space learning
is fast and most of the labels falls on this space, and (iii) label noise falls
on the nuisance space and impedes optimization/generalization.
****

Learning to Remember from a Multi-Task Teacher
Yuwen Xiong,Mengye Ren,Raquel Urtasun
Recent studies on catastrophic forgetting during sequential learning typically f
ocus on fixing the accuracy of the predictions for a previously learned task. In
 this paper we argue that the outputs of neural networks are subject to rapid ch

anges when learning a new data distribution, and networks that appear to "forget " everything still contain useful representation towards previous tasks. We thus propose to enforce the output accuracy to stay the same, we should aim to reduce the effect of catastrophic forgetting on the representation level, as the output layer can be quickly recovered later with a small number of examples. Towards this goal, we propose an experimental setup that measures the amount of representational forgetting, and develop a novel meta-learning algorithm to overcome this issue. The proposed meta-learner produces weight updates of a sequential learning network, mimicking a multi-task teacher network's representation. We show that our meta-learner can improve its learned representations on new tasks, while maintaining a good representation for old tasks.
****

Gradient ℓ_1 Regularization for Quantization Robustness
Milad Alizadeh,Arash Behboodi,Mart van Baalen,Christos Louizos,Tijmen Blankevoort,Max Welling
We analyze the effect of quantizing weights and activations of neural networks on their loss and derive a simple regularization scheme that improves robustness against post-training quantization. By training quantization-ready networks, our approach enables storing a single set of weights that can be quantized on-demand to different bit-widths as energy and memory requirements of the application change. Unlike quantization-aware training using the straight-through estimator that only targets a specific bit-width and requires access to training data and pipeline, our regularization-based method paves the way for ``on the fly'' post-training quantization to various bit-widths. We show that by modeling quantization as a ℓ_∞-bounded perturbation, the first-order term in the loss expansion can be regularized using the ℓ_1-norm of gradients. We experimentally validate our method on different vision architectures on CIFAR-10 and ImageNet datasets and show that the regularization of a neural network using our method improves robustness against quantization noise.
****

Coloring graph neural networks for node disambiguation
George Dasoulas,Ludovic Dos Santos,Kevin Scaman,Aladin Virmaux
In this paper, we show that a simple coloring scheme can improve, both theoretically and empirically, the expressive power of Message Passing Neural Networks (MPNNs). More specifically, we introduce a graph neural network called Colored Local Iterative Procedure (CLIP) that uses colors to disambiguate identical node attributes, and show that this representation is a universal approximator of continuous functions on graphs with node attributes. Our method relies on separability, a key topological characteristic that allows to extend well-chosen neural networks into universal representations. Finally, we show experimentally that CLIP is capable of capturing structural characteristics that traditional MPNNs fail to distinguish, while being state-of-the-art on benchmark graph classification datasets.
****

Spectral Embedding of Regularized Block Models
Nathan De Lara,Thomas Bonald
Spectral embedding is a popular technique for the representation of graph data. Several regularization techniques have been proposed to improve the quality of the embedding with respect to downstream tasks like clustering. In this paper, we explain on a simple block model the impact of the complete graph regularization, whereby a constant is added to all entries of the adjacency matrix. Specifically, we show that the regularization forces the spectral embedding to focus on the largest blocks, making the representation less sensitive to noise or outliers. We illustrate these results on both on both synthetic and real data, showing how regularization improves standard clustering scores.
****

On Federated Learning of Deep Networks from Non-IID Data: Parameter Divergence and the Effects of Hyperparametric Methods
Heejae Kim,Taewoo Kim,Chan-Hyun Youn
Federated learning, where a global model is trained by iterative parameter avera

ging of locally-computed updates, is a promising approach for distributed traini
ng of deep networks; it provides high communication-efficiency and privacy-prese
rvability, which allows to fit well into decentralized data environments, e.g.,
mobile-cloud ecosystems. However, despite the advantages, the federated learning
-based methods still have a challenge in dealing with non-IID training data of l
ocal devices (i.e., learners). In this regard, we study the effects of a variety
 of hyperparametric conditions under the non-IID environments, to answer importa
nt concerns in practical implementations: (i) We first investigate parameter div
ergence of local updates to explain performance degradation from non-IID data. T
he origin of the parameter divergence is also found both empirically and theoret
ically. (ii) We then revisit the effects of optimizers, network depth/width, and
 regularization techniques; our observations show that the well-known advantages
 of the hyperparameter optimization strategies could rather yield diminishing re
turns with non-IID data. (iii) We finally provide the reasons of the failure cas
es in a categorized way, mainly based on metrics of the parameter divergence.
****

Improved Detection of Adversarial Attacks via Penetration Distortion Maximizatio
n

Shai Rozenberg,Gal Elidan,Ran El-Yaniv

This paper is concerned with the defense of deep models against adversarial at-
tacks. We develop an adversarial detection method, which is inspired by the cer-
tificate defense approach, and captures the idea of separating class clusters in
 the
embedding space so as to increase the margin. The resulting defense is intuitive
,
effective, scalable and can be integrated into any given neural classification m
odel.
Our method demonstrates state-of-the-art detection performance under all threat
models.
****

Barcodes as summary of objective functions' topology

Serguei Barannikov,Alexander Korotin,Dmitry Oganesyan,Daniil Emtsev,Evgeny Burna
ev

We apply canonical forms of gradient complexes (barcodes) to explore neural netw
orks loss surfaces. We present an algorithm for calculations of the objective fu
nction's barcodes of minima. Our experiments confirm two principal observations
: (1) the barcodes of minima are located in a small lower part of the range of v
alues of objective function and (2) increase of the neural network's depth bring
s down the minima's barcodes. This has natural implications for the neural netwo
rk learning and the ability to generalize.
****

Toward Evaluating Robustness of Deep Reinforcement Learning with Continuous Cont
rol

Tsui-Wei Weng,Krishnamurthy (Dj) Dvijotham*,Jonathan Uesato*,Kai Xiao*,Sven Gowa
l*,Robert Stanforth*,Pushmeet Kohli

Deep reinforcement learning has achieved great success in many previously diffic
ult reinforcement learning tasks, yet recent studies show that deep RL agents ar
e also unavoidably susceptible to adversarial perturbations, similar to deep neu
ral networks in classification tasks. Prior works mostly focus on model-free adv
ersarial attacks and agents with discrete actions. In this work, we study the pr
oblem of continuous control agents in deep RL with adversarial attacks and propo
se the first two-step algorithm based on learned model dynamics. Extensive exper
iments on various MuJoCo domains (Cartpole, Fish, Walker, Humanoid) demonstrate
that our proposed framework is much more effective and efficient than model-free
 based attacks baselines in degrading agent performance as well as driving agent
s to unsafe states.
****

LEARNING TO IMPUTE: A GENERAL FRAMEWORK FOR SEMI-SUPERVISED LEARNING

Wei-Hong Li,Chuan-Sheng Foo,Hakan Bilen

Recent semi-supervised learning methods have shown to achieve comparable results

to their supervised counterparts while using only a small portion of labels in image classification tasks thanks to their regularization strategies. In this paper, we take a more direct approach for semi-supervised learning and propose learning to impute the labels of unlabeled samples such that a network achieves better generalization when it is trained on these labels. We pose the problem in a learning-to-learn formulation which can easily be incorporated to the state-of-the-art semi-supervised techniques and boost their performance especially when the labels are limited. We demonstrate that our method is applicable to both classification and regression problems including image classification and facial landmark detection tasks.
****

Geometry-aware Generation of Adversarial and Cooperative Point Clouds
Yuxin Wen,Jiehong Lin,Ke Chen,Kui Jia
Recent studies show that machine learning models are vulnerable to adversarial examples. In 2D image domain, these examples are obtained by adding imperceptible noises to natural images. This paper studies adversarial generation of point clouds by learning to deform those approximating object surfaces of certain categories. As 2D manifolds embedded in the 3D Euclidean space, object surfaces enjoy the general properties of smoothness and fairness. We thus argue that in order to achieve imperceptible surface shape deformations, adversarial point clouds should have the same properties with similar degrees of smoothness/fairness to the benign ones, while being close to the benign ones as well when measured under certain distance metrics of point clouds. To this end, we propose a novel loss function to account for imperceptible, geometry-aware deformations of point clouds, and use the proposed loss in an adversarial objective to attack representative models of point set classifiers. Experiments show that our proposed method achieves stronger attacks than existing methods, without introduction of noticeable outliers and surface irregularities. In this work, we also investigate an opposite direction that learns to deform point clouds of object surfaces in the same geometry-aware, but cooperative manner. Cooperatively generated point clouds are more favored by machine learning models in terms of improved classification confidence or accuracy. We present experiments verifying that our proposed objective succeeds in learning cooperative shape deformations.
****

Crafting Data-free Universal Adversaries with Dilate Loss
Deepak Babu Sam,ABINAYA K,Sudharsan K A,Venkatesh Babu RADHAKRISHNAN
We introduce a method to create Universal Adversarial Perturbations (UAP) for a given CNN in a data-free manner. Data-free approaches suite scenarios where the original training data is unavailable for crafting adversaries. We show that the adversary generation with full training data can be approximated to a formulation without data. This is realized through a sequential optimization of the adversarial perturbation with the proposed dilate loss. Dilate loss basically maximizes the Euclidean norm of the output before nonlinearity at any layer. By doing so, the perturbation constrains the ReLU activation function at every layer to act roughly linear for data points and thus eliminate the dependency on data for crafting UAPs. Extensive experiments demonstrate that our method not only has theoretical support, but achieves higher fooling rate than the existing data-free work. Furthermore, we evidence improvement in limited data cases as well.
****

Efficient Bi-Directional Verification of ReLU Networks via Quadratic Programming
Aleksei Kuvshinov,Stephan Guennemann
Neural networks are known to be sensitive to adversarial perturbations. To investigate this undesired behavior we consider the problem of computing the distance to the decision boundary (DtDB) from a given sample for a deep NN classifier. In this work we present an iterative procedure where in each step we solve a convex quadratic programming (QP) task. Solving the single initial QP already results in a lower bound on the DtDB and can be used as a robustness certificate of the classifier around a given sample. In contrast to currently known approaches our method also provides upper bounds used as a measure of quality for the certificate. We show that our approach provides better or competitive results in compar

ison with a wide range of existing techniques.
****
Improving Sample Efficiency in Model-Free Reinforcement Learning from Images
Denis Yarats,Amy Zhang,Ilya Kostrikov,Brandon Amos,Joelle Pineau,Rob Fergus
Training an agent to solve control tasks directly from high-dimensional images with model-free reinforcement learning (RL) has proven difficult. The agent needs to learn a latent representation together with a control policy to perform the task. Fitting a high-capacity encoder using a scarce reward signal is not only extremely sample inefficient, but also prone to suboptimal convergence. Two ways to improve sample efficiency are to learn a good feature representation and use off-policy algorithms. We dissect various approaches of learning good latent features, and conclude that the image reconstruction loss is the essential ingredient that enables efficient and stable representation learning in image-based RL. Following these findings, we devise an off-policy actor-critic algorithm with an auxiliary decoder that trains end-to-end and matches state-of-the-art performance across both model-free and model-based algorithms on many challenging control tasks. We release our code to encourage future research on image-based RL.
****
Improving Exploration of Deep Reinforcement Learning using Planning for Policy Search
Jakob J. Hollenstein,Erwan Renaudo,Justus Piater
Most Deep Reinforcement Learning methods perform local search and
therefore are prone to get stuck on non-optimal
solutions. Furthermore, in simulation based training, such as
domain-randomized simulation training, the availability of a simulation
model is not exploited, which potentially decreases
efficiency. To overcome issues of local search and exploit
access to simulation models, we propose the use of kino-dynamic
planning methods as part of a model-based reinforcement learning
method and to learn in an off-policy fashion from solved planning
instances. We show that, even on a simple toy domain, D-RL
methods (DDPG, PPO, SAC) are not immune to local optima and
require additional exploration mechanisms. We show that our
planning method exhibits a better state space coverage, collects
data that allows for better policies than D-RL methods without
additional exploration mechanisms and that starting from the
planner data and performing additional training results in as
good as or better policies than vanilla D-RL methods, while also
creating data that is more fit for re-use in modified tasks.

****
A Theoretical Analysis of Deep Q-Learning
Zhuoran Yang,Yuchen Xie,Zhaoran Wang
Despite the great empirical success of deep reinforcement learning, its theoretical foundation is less well understood. In this work, we make the first attempt to theoretically understand the deep Q-network (DQN) algorithm (Mnih et al., 2015) from both algorithmic and statistical perspectives. In specific, we focus on a slight simplification of DQN that fully captures its key features. Under mild assumptions, we establish the algorithmic and statistical rates of convergence for the action-value functions of the iterative policy sequence obtained by DQN. In particular, the statistical error characterizes the bias and variance that arise from approximating the action-value function using deep neural network, while the algorithmic error converges to zero at a geometric rate. As a byproduct, our analysis provides justifications for the techniques of experience replay and target network, which are crucial to the empirical success of DQN. Furthermore, as a simple extension of DQN, we propose the Minimax-DQN algorithm for zero-sum Markov game with two players, which is deferred to the appendix due to space limitations.
****
Decentralized Deep Learning with Arbitrary Communication Compression

Anastasia Koloskova*,Tao Lin*,Sebastian U Stich,Martin Jaggi
Decentralized training of deep learning models is a key element for enabling dat a privacy and on-device learning over networks, as well as for efficient scaling to large compute clusters. As current approaches are limited by network bandwid th, we propose the use of communication compression in the decentralized trainin g context. We show that Choco-SGD achieves linear speedup in the number of worke rs for arbitrary high compression ratios on general non-convex functions, and no n-IID training data. We demonstrate the practical performance of the algorithm in two key scenarios: the training of deep learning models (i) over decentralize d user devices, connected by a peer-to-peer network and (ii) in a datacenter.
****

Can I Trust the Explainer? Verifying Post-Hoc Explanatory Methods
Oana-Maria Camburu*,Eleonora Giunchiglia*,Jakob Foerster,Thomas Lukasiewicz,Phil Blunsom
For AI systems to garner widespread public acceptance, we must develop methods c apable of explaining the decisions of black-box models such as neural networks. In this work, we identify two issues of current explanatory methods. First, we s how that two prevalent perspectives on explanations—feature-additivity and featu re-selection—lead to fundamentally different instance-wise explanations. In the literature, explainers from different perspectives are currently being directly compared, despite their distinct explanation goals. The second issue is that cur rent post-hoc explainers have only been thoroughly validated on simple models, s uch as linear regression, and, when applied to real-world neural networks, expla iners are commonly evaluated under the assumption that the learned models behave reasonably. However, neural networks often rely on unreasonable correlations, e ven when producing correct decisions. We introduce a verification framework for explanatory methods under the feature-selection perspective. Our framework is ba sed on a non-trivial neural network architecture trained on a real-world task, a nd for which we are able to provide guarantees on its inner workings. We validat e the efficacy of our evaluation by showing the failure modes of current explain ers. We aim for this framework to provide a publicly available,1 off-the-shelf e valuation when the feature-selection perspective on explanations is needed.
****

D3PG: Deep Differentiable Deterministic Policy Gradients
Tao Du,Yunfei Li,Jie Xu,Andrew Spielberg,Kui Wu,Daniela Rus,Wojciech Matusik
Over the last decade, two competing control strategies have emerged for solving complex control tasks with high efficacy. Model-based control algorithms, such a s model-predictive control (MPC) and trajectory optimization, peer into the grad ients of underlying system dynamics in order to solve control tasks with high sa mple efficiency. However, like all gradient-based numerical optimization method s,model-based control methods are sensitive to intializations and are prone to b ecoming trapped in local minima. Deep reinforcement learning (DRL), on the other hand, can somewhat alleviate these issues by exploring the solution space throu gh sampling – at the expense of computational cost. In this paper, we present a hybrid method that combines the best aspects of gradient-based methods and DRL. We base our algorithm on the deep deterministic policy gradients (DDPG) algorith m and propose a simple modification that uses true gradients from a differentiab le physical simulator to increase the convergence rate of both the actor and the critic. We demonstrate our algorithm on seven 2D robot control tasks, with the most complex one being a differentiable half cheetah with hard contact constrai nts. Empirical results show that our method boosts the performance of DDPGwithou t sacrificing its robustness to local minima.
****

Deep Ensembles: A Loss Landscape Perspective
Stanislav Fort,Clara Huiyi Hu,Balaji Lakshminarayanan
Deep ensembles have been empirically shown to be a promising approach for improv ing accuracy, uncertainty and out-of-distribution robustness of deep learning m odels. While deep ensembles were theoretically motivated by the bootstrap, non-b ootstrap ensembles trained with just random initialization also perform well in practice, which suggests that there could be other explanations for why deep ens

embles work well. Bayesian neural networks, which learn distributions over the p
arameters of the network, are theoretically well-motivated by Bayesian principle
s, but do not perform as well as deep ensembles in practice, particularly under
dataset shift. One possible explanation for this gap between theory and practice
 is that popular scalable approximate Bayesian methods tend to focus on a single
 mode, whereas deep ensembles tend to explore diverse modes in function space. W
e investigate this hypothesis by building on recent work on understanding the lo
ss landscape of neural networks and adding our own exploration to measure the si
milarity of functions in the space of predictions. Our results show that random
initializations explore entirely different modes, while functions along an optim
ization trajectory or sampled from the subspace thereof cluster within a single
mode predictions-wise, while often deviating significantly in the weight space.
We demonstrate that while low-loss connectors between modes exist, they are not
connected in the space of predictions. Developing the concept of the diversity--
accuracy plane, we show that the decorrelation power of random initializations i
s unmatched by popular subspace sampling methods.
****

A Finite-Time Analysis of Q-Learning with Neural Network Function Approximation
Pan Xu,Quanquan Gu
Q-learning with neural network function approximation (neural Q-learning for sho
rt) is among the most prevalent deep reinforcement learning algorithms. Despite
its empirical success, the non-asymptotic convergence rate of neural Q-learning
remains virtually unknown. In this paper, we present a finite-time analysis of a
 neural Q-learning algorithm, where the data are generated from a Markov decisio
n process and the action-value function is approximated by a deep ReLU neural ne
twork. We prove that neural Q-learning finds the optimal policy with $O(1/T)$ co
nvergence rate if the neural function approximator is sufficiently overparameter
ized, where T is the number of iterations. To our best knowledge, our result i
s the first finite-time analysis of neural Q-learning under non-i.i.d. data assu
mption.
****

MULTI-STAGE INFLUENCE FUNCTION
Hongge Chen,Si Si,Yang Li,Ciprian Chelba,Sanjiv Kumar,Duane Boning,Cho-Jui Hsieh
Multi-stage training and knowledge transfer from a large-scale pretrain task to
various fine-tune end tasks have revolutionized natural language processing (NLP
) and computer vision (CV), with state-of-the-art performances constantly being
improved. In this paper, we develop a multi-stage influence function score to tr
ack predictions from a finetune model all the way back to the pretrain data. Wit
h this score, we can identify the pretrain examples in the pretrain task that co
ntribute most to a prediction in the fine-tune task. The proposed multi-stage in
fluence function generalizes the original influence function for a single model
in Koh et al 2017, thereby enabling influence computation through both pretrain
and fine-tune models. We test our proposed method in various experiments to show
 its effectiveness and potential applications.
****

Impact of the latent space on the ability of GANs to fit the distribution
Thomas Pinetz,Daniel Soukup,Thomas Pock
The goal of generative models is to model the underlying data distribution of a
sample based dataset. Our intuition is that an accurate model should in principl
e
also include the sample based dataset as part of its induced probability distrib
ution.
To investigate this, we look at fully trained generative models using the Genera
tive
Adversarial Networks (GAN) framework and analyze the resulting generator
on its ability to memorize the dataset. Further, we show that the size of the in
itial
latent space is paramount to allow for an accurate reconstruction of the trainin
g
data. This gives us a link to compression theory, where Autoencoders (AE) are

used to lower bound the reconstruction capabilities of our generative model. Her
e,
we observe similar results to the perception-distortion tradeoff (Blau & Michael
i
(2018)). Given a small latent space, the AE produces low quality and the GAN
produces high quality outputs from a perceptual viewpoint. In contrast, the dist
ortion
error is smaller for the AE. By increasing the dimensionality of the latent
space the distortion decreases for both models, but the perceptual quality only
increases for the AE.
****

Training Generative Adversarial Networks from Incomplete Observations using Fact
orised Discriminators
Daniel Stoller,Sebastian Ewert,Simon Dixon
Generative adversarial networks (GANs) have shown great success in applications
such as image generation and inpainting.
However, they typically require large datasets, which are often not available, e
specially in the context of prediction tasks such as image segmentation that req
uire labels. Therefore, methods such as the CycleGAN use more easily available u
nlabelled data, but do not offer a way to leverage additional labelled data for
improved performance. To address this shortcoming, we show how to factorise the
joint data distribution into a set of lower-dimensional distributions along with
 their dependencies. This allows splitting the discriminator in a GAN into multi
ple "sub-discriminators" that can be independently trained from incomplete obser
vations. Their outputs can be combined to estimate the density ratio between the
 joint real and the generator distribution, which enables training generators as
 in the original GAN framework. We apply our method to image generation, image s
egmentation and audio source separation, and obtain improved performance over a
standard GAN when additional incomplete training examples are available. For the
 Cityscapes segmentation task in particular, our method also improves accuracy b
y an absolute 14.9% over CycleGAN while using only 25 additional paired examples
.
****

Combining Q-Learning and Search with Amortized Value Estimates
Jessica B. Hamrick,Victor Bapst,Alvaro Sanchez-Gonzalez,Tobias Pfaff,Theophane W
eber,Lars Buesing,Peter W. Battaglia
We introduce "Search with Amortized Value Estimates" (SAVE), an approach for com
bining model-free Q-learning with model-based Monte-Carlo Tree Search (MCTS). In
 SAVE, a learned prior over state-action values is used to guide MCTS, which est
imates an improved set of state-action values. The new Q-estimates are then used
 in combination with real experience to update the prior. This effectively amort
izes the value computation performed by MCTS, resulting in a cooperative relatio
nship between model-free learning and model-based search. SAVE can be implemente
d on top of any Q-learning agent with access to a model, which we demonstrate by
 incorporating it into agents that perform challenging physical reasoning tasks
and Atari. SAVE consistently achieves higher rewards with fewer training steps,
and---in contrast to typical model-based search approaches---yields strong perfo
rmance with very small search budgets. By combining real experience with informa
tion computed during search, SAVE demonstrates that it is possible to improve on
 both the performance of model-free learning and the computational cost of plann
ing.
****

Infinite-Horizon Differentiable Model Predictive Control
Sebastian East,Marco Gallieri,Jonathan Masci,Jan Koutnik,Mark Cannon
This paper proposes a differentiable linear quadratic Model Predictive Control (
MPC) framework for safe imitation learning. The infinite-horizon cost is enforce
d using a terminal cost function obtained from the discrete-time algebraic Ricca
ti equation (DARE), so that the learned controller can be proven to be stabilizi
ng in closed-loop. A central contribution is the derivation of the analytical de
rivative of the solution of the DARE, thereby allowing the use of differentiatio

n-based learning methods. A further contribution is the structure of the MPC opt imization problem: an augmented Lagrangian method ensures that the MPC optimizat ion is feasible throughout training whilst enforcing hard constraints on state a nd input, and a pre-stabilizing controller ensures that the MPC solution and der ivatives are accurate at each iteration. The learning capabilities of the framew ork are demonstrated in a set of numerical studies.
****

Anchor & Transform: Learning Sparse Representations of Discrete Objects
Paul Pu Liang,Manzil Zaheer,Yuan Wang,Amr Ahmed
Learning continuous representations of discrete objects such as text, users, and items lies at the heart of many applications including text and user modeling. Unfortunately, traditional methods that embed all objects do not scale to large vocabulary sizes and embedding dimensions. In this paper, we propose a general m ethod, Anchor & Transform (ANT) that learns sparse representations of discrete o bjects by jointly learning a small set of anchor embeddings and a sparse transfo rmation from anchor objects to all objects. ANT is scalable, flexible, end-to-en d trainable, and allows the user to easily incorporate domain knowledge about ob ject relationships (e.g. WordNet, co-occurrence, item clusters). ANT also recove rs several task-specific baselines under certain structural assumptions on the a nchors and transformation matrices. On text classification and language modeling benchmarks, ANT demonstrates stronger performance with fewer parameters as comp ared to existing vocabulary selection and embedding compression baselines.
****

Emergence of Collective Policies Inside Simulations with Biased Representations
Jooyeon Kim,Alice Oh
We consider a setting where biases are involved when agents internalise an envir onment. Agents have different biases, all of which resulting in imperfect evide nce collected for taking optimal actions. Throughout the interactions, each age nt asynchronously internalises their own predictive model of the environment and forms a virtual simulation within which the agent plays trials of the episodes in entirety. In this research, we focus on developing a collective policy traine d solely inside agents' simulations, which can then be transferred to the real-w orld environment. The key idea is to let agents imagine together; make them take turns to host virtual episodes within which all agents participate and interact with their own biased representations. Since agents' biases vary, the collectiv e policy developed while sequentially visiting the internal simulations compleme nt one another's shortcomings. In our experiment, the collective policies consis tently achieve significantly higher returns than the best individually trained p olicies.
****

Projection-Based Constrained Policy Optimization
Tsung-Yen Yang,Justinian Rosca,Karthik Narasimhan,Peter J. Ramadge
We consider the problem of learning control policies that optimize a reward func tion while satisfying constraints due to considerations of safety, fairness, or other costs. We propose a new algorithm - Projection-Based Constrained Policy Op timization (PCPO), an iterative method for optimizing policies in a two-step pro cess - the first step performs an unconstrained update while the second step rec onciles the constraint violation by projecting the policy back onto the constrai nt set. We theoretically analyze PCPO and provide a lower bound on reward improv ement, as well as an upper bound on constraint violation for each policy update. We further characterize the convergence of PCPO with projection based on two di fferent metrics - L2 norm and Kullback-Leibler divergence. Our empirical results over several control tasks demonstrate that our algorithm achieves superior per formance, averaging more than 3.5 times less constraint violation and around 15% higher reward compared to state-of-the-art methods.
****

Maximum Likelihood Constraint Inference for Inverse Reinforcement Learning
Dexter R.R. Scobee,S. Shankar Sastry
While most approaches to the problem of Inverse Reinforcement Learning (IRL) foc us on estimating a reward function that best explains an expert agent’s policy o

r demonstrated behavior on a control task, it is often the case that such behavi
or is more succinctly represented by a simple reward combined with a set of hard
 constraints. In this setting, the agent is attempting to maximize cumulative re
wards subject to these given constraints on their behavior. We reformulate the p
roblem of IRL on Markov Decision Processes (MDPs) such that, given a nominal mod
el of the environment and a nominal reward function, we seek to estimate state,
action, and feature constraints in the environment that motivate an agent’s beha
vior. Our approach is based on the Maximum Entropy IRL framework, which allows u
s to reason about the likelihood of an expert agent’s demonstrations given our k
nowledge of an MDP. Using our method, we can infer which constraints can be adde
d to the MDP to most increase the likelihood of observing these demonstrations.
We present an algorithm which iteratively infers the Maximum Likelihood Constrai
nt to best explain observed behavior, and we evaluate its efficacy using both si
mulated behavior and recorded data of humans navigating around an obstacle.
****

Towards Effective 2-bit Quantization: Pareto-optimal Bit Allocation for Deep CNN
s Compression

Zhe Wang,Jie Lin,Mohamed M. Sabry Aly,Sean I Young,Vijay Chandrasekhar,Bernd Gir
od

State-of-the-art quantization methods can compress deep neural networks down to
4 bits without losing accuracy. However, when it comes to 2 bits, the performanc
e drop is still noticeable. One problem in these methods is that they assign equ
al bit rate to quantize weights and activations in all layers, which is not reas
onable in the case of high rate compression (such as 2-bit quantization), as som
e of layers in deep neural networks are sensitive to quantization and performing
 coarse quantization on these layers can hurt the accuracy. In this paper, we ad
dress an important problem of how to optimize the bit allocation of weights and
activations for deep CNNs compression. We first explore the additivity of output
 error caused by quantization and find that additivity property holds for deep n
eural networks which are continuously differentiable in the layers. Based on thi
s observation, we formulate the optimal bit allocation problem of weights and ac
tivations in a joint framework and propose a very efficient method to solve the
optimization problem via Lagrangian Formulation. Our method obtains excellent re
sults on deep neural networks. It can compress deep CNN ResNet-50 down to 2 bits
 with only 0.7% accuracy loss. To the best our knowledge, this is the first pape
r that reports 2-bit results on deep CNNs without hurting the accuracy.
****

You Only Train Once: Loss-Conditional Training of Deep Networks

Alexey Dosovitskiy,Josip Djolonga

In many machine learning problems, loss functions are weighted sums of several t
erms. A typical approach to dealing with these is to train multiple separate mod
els with different selections of weights and then either choose the best one acc
ording to some criterion or keep multiple models if it is desirable to maintain
a diverse set of solutions. This is inefficient both at training and at inferenc
e time. We propose a method that allows replacing multiple models trained on one
 loss function each by a single model trained on a distribution of losses. At te
st time a model trained this way can be conditioned to generate outputs correspo
nding to any loss from the training distribution of losses. We demonstrate this
approach on three tasks with parametrized losses: beta-VAE, learned image compre
ssion, and fast style transfer.
****

Meta-Learning Acquisition Functions for Transfer Learning in Bayesian Optimizati
on

Michael Volpp,Lukas P. Fröhlich,Kirsten Fischer,Andreas Doerr,Stefan Falkner,Fra
nk Hutter,Christian Daniel

Transferring knowledge across tasks to improve data-efficiency is one of the ope
n key challenges in the field of global black-box optimization. Readily availabl
e algorithms are typically designed to be universal optimizers and, therefore, o
ften suboptimal for specific tasks. We propose a novel transfer learning method
to obtain customized optimizers within the well-established framework of Bayesia

n optimization, allowing our algorithm to utilize the proven generalization capa
bilities of Gaussian processes. Using reinforcement learning to meta-train an ac
quisition function (AF) on a set of related tasks, the proposed method learns to
 extract implicit structural information and to exploit it for improved data-eff
iciency. We present experiments on a simulation-to-real transfer task as well as
 on several synthetic functions and on two hyperparameter search problems. The r
esults show that our algorithm (1) automatically identifies structural propertie
s of objective functions from available source tasks or simulations, (2) perform
s favourably in settings with both scarse and abundant source data, and (3) fall
s back to the performance level of general AFs if no particular structure is pre
sent.
****

Using Explainabilty to Detect Adversarial Attacks
Ohad Amosy and Gal Chechik
Deep learning models are often sensitive to adversarial attacks, where carefully
-designed input samples can cause the system to produce incorrect decisions. Her
e we focus on the problem of detecting attacks, rather than robust classificatio
n, since detecting that an attack occurs may be even more important than avoidin
g misclassification. We build on advances in explainability, where activity-map-
like explanations are used to justify and validate decisions, by highlighting fe
atures that are involved with a classification decision. The key observation is
that it is hard to create explanations for incorrect decisions. We propose EXAI
D, a novel attack-detection approach, which uses model explainability to identif
y images whose explanations are inconsistent with the predicted class. Specifica
lly, we use SHAP, which uses Shapley values in the space of the input image, to
identify which input features contribute to a class decision. Interestingly, thi
s approach does not require to modify the attacked model, and it can be applied
without modelling a specific attack. It can therefore be applied successfully to
 detect unfamiliar attacks, that were unknown at the time the detection model wa
s designed. We evaluate EXAID on two benchmark datasets CIFAR-10 and SVHN, and a
gainst three leading attack techniques, FGSM, PGD and C&W. We find that EXAID im
proves over the SoTA detection methods by a large margin across a wide range of
noise levels, improving detection from 70% to over 90% for small perturbations.
****

Feature Selection using Stochastic Gates
Yutaro Yamada,Ofir Lindenbaum,Sahand Negahban,Yuval Kluger
Feature selection problems have been extensively studied in the setting of
linear estimation, for instance LASSO, but less emphasis has been placed on
feature selection for non-linear functions. In this study, we propose a method
for feature selection in high-dimensional non-linear function estimation
problems. The new procedure is based on directly penalizing the ℓ_0 norm of
features, or the count of the number of selected features. Our ℓ_0 based re
gularization relies on a continuous relaxation of the Bernoulli distribution, wh
ich
allows our model to learn the parameters of the approximate Bernoulli
distributions via gradient descent. The proposed framework simultaneously learns
a non-linear regression or classification function while selecting a small
subset of features. We provide an information-theoretic justification for
incorporating Bernoulli distribution into our approach. Furthermore, we evaluate
our method using synthetic and real-life data and demonstrate that our approach
outperforms other embedded methods in terms of predictive performance and featur
e selection.
****

SpectroBank: A filter-bank convolutional layer for CNN-based audio applications
Helena Peic Tukuljac,Benjamin Ricaud,Nicolas Aspert,Pierre Vandergheynst
We propose and investigate the design of a new convolutional layer where kernels
 are parameterized functions. This layer aims at being the input layer of convol
utional neural networks for audio applications. The kernels are defined as funct
ions having a band-pass filter shape, with a limited number of trainable paramet
ers. We show that networks having such an input layer can achieve state-of-the-a

rt accuracy on several audio classification tasks. This approach, while reducing the number of weights to be trained along with network training time, enables larger kernel sizes, an advantage for audio applications. Furthermore, the learned filters bring additional interpretability and a better understanding of the data properties exploited by the network.
****
Testing For Typicality with Respect to an Ensemble of Learned Distributions
Forrest Laine,Claire Tomlin
Good methods of performing anomaly detection on high-dimensional data sets are needed, since algorithms which are trained on data are only expected to perform well on data that is similar to the training data. There are theoretical results on the ability to detect if a population of data is likely to come from a known base distribution, which is known as the goodness-of-fit problem, but those results require knowing a model of the base distribution. The ability to correctly reject anomalous data hinges on the accuracy of the model of the base distribution. For high dimensional data, learning an accurate-enough model of the base distribution such that anomaly detection works reliably is very challenging, as many researchers have noted in recent years. Existing methods for the goodness-of-fit problem do not account for the fact that a model of the base distribution is learned. To address that gap, we offer a theoretically motivated approach to account for the density learning procedure. In particular, we propose training an ensemble of density models, considering data to be anomalous if the data is anomalous with respect to any member of the ensemble. We provide a theoretical justification for this approach, proving first that a test on typicality is a valid approach to the goodness-of-fit problem, and then proving that for a correctly constructed ensemble of models, the intersection of typical sets of the models lies in the interior of the typical set of the base distribution. We present our method in the context of an example on synthetic data in which the effects we consider can easily be seen.
****
GraphSAINT: Graph Sampling Based Inductive Learning Method
Hanqing Zeng,Hongkuan Zhou,Ajitesh Srivastava,Rajgopal Kannan,Viktor Prasanna
Graph Convolutional Networks (GCNs) are powerful models for learning representations of attributed graphs. To scale GCNs to large graphs, state-of-the-art methods use various layer sampling techniques to alleviate the "neighbor explosion" problem during minibatch training. We propose GraphSAINT, a graph sampling based inductive learning method that improves training efficiency and accuracy in a fundamentally different way. By changing perspective, GraphSAINT constructs mini batches by sampling the training graph, rather than the nodes or edges across GCN layers. Each iteration, a complete GCN is built from the properly sampled subgraph. Thus, we ensure fixed number of well-connected nodes in all layers. We further propose normalization technique to eliminate bias, and sampling algorithms for variance reduction. Importantly, we can decouple the sampling from the forward and backward propagation, and extend GraphSAINT with many architecture variants (e.g., graph attention, jumping connection). GraphSAINT demonstrates superior performance in both accuracy and training time on five large graphs, and achieves new state-of-the-art F1 scores for PPI (0.995) and Reddit (0.970).
****
Adversarial Filters of Dataset Biases
Ronan Le Bras,Swabha Swayamdipta,Chandra Bhagavatula,Rowan Zellers,Matthew Peters,Ashish Sabharwal,Yejin Choi

Large-scale benchmark datasets have been among the major driving forces in AI, supporting training of models and measuring their progress. The key assumption is that these benchmarks are realistic approximations of the target tasks in the real world. However, while machine performance on these benchmarks advances rapidly --- often surpassing human performance --- it still struggles on the target tasks in the wild. This raises an important question: whether the surreal high performance on existing benchmarks are inflated due to spurious biases in them, and if so, how we can effectively revise these benchmarks to better simulate more realistic problem distributions in the real world.
In this paper, we posit that while the real world problems consist of a great deal of long-tail problems, existing benchmarks are overly populated with a great deal of similar (thus non-tail) problems, which in turn, leads to a major overestimation of true AI performance. To address this challenge, we present a novel framework of Adversarial Filters to investigate model-based reduction of dataset biases. We discuss that the optimum bias reduction via AFOptimum is intractable, thus propose AFLite, an iterative greedy algorithm that adversarially filters out data points to identify a reduced dataset with more realistic problem distributions and considerably less spurious biases.
AFLite is lightweight and can in principle be applied to any task and dataset. We apply it to popular benchmarks that are practically solved --- ImageNet and Natural Language Inference (SNLI, MNLI, QNLI) --- and present filtered counterparts as new challenge datasets where the model performance drops considerably (e.g., from 84% to 24% for ImageNet and from 92% to 62% for SNLI), while human performance remains high. An extensive suite of analysis demonstrates that AFLite effectively reduces measurable dataset biases in both the synthetic and real datasets. Finally, we introduce new measures of dataset biases based on K-nearest-neighbors to help guide future research on dataset developments and bias reduction.
****

Value-Driven Hindsight Modelling
Arthur Guez,Fabio Viola,Theophane Weber,Lars Buesing,Steven Kapturowski,Doina Precup,David Silver,Nicolas Heess
Value estimation is a critical component of the reinforcement learning (RL) paradigm. The question of how to effectively learn predictors for value from data is one of the major problems studied by the RL community, and different approaches exploit structure in the problem domain in different ways. Model learning can make use of the rich transition structure present in sequences of observations, but this approach is usually not sensitive to the reward function. In contrast, model-free methods directly leverage the quantity of interest from the future but have to compose with a potentially weak scalar signal (an estimate of the return). In this paper we develop an approach for representation learning in RL that sits in between these two extremes: we propose to learn what to model in a way that can directly help value prediction. To this end we determine which features of the future trajectory provide useful information to predict the associated return. This provides us with tractable prediction targets that are directly relevant for a task, and can thus accelerate learning of the value function. The idea can be understood as reasoning, in hindsight, about which aspects of the future observations could help past value prediction. We show how this can help dramatically even in simple policy evaluation settings. We then test our approach at scale in challenging domains, including on 57 Atari 2600 games.
****

Learning Neural Causal Models from Unknown Interventions
Nan Rosemary Ke,Olexa Bilaniuk,Anirudh Goyal,Stephan Bauer,Hugol Larochelle,Chris Pal,Yoshua Bengio
Meta-learning over a set of distributions can be interpreted as learning different types of parameters corresponding to short-term vs long-term aspects of the mechanisms underlying the generation of data. These are respectively captured by quickly-changing \textit{parameters} and slowly-changing \textit{meta-parameters}. We present a new framework for meta-learning causal models where the relationship between each variable and its parents is modeled by a neural network, modulated by structural meta-parameters which capture the overall topology of a direc

ted graphical model. Our approach avoids a discrete search over models in favour of a continuous optimization procedure. We study a setting where interventional distributions are induced as a result of a random intervention on a single unknown variable of an unknown ground truth causal model, and the observations arising after such an intervention constitute one meta-example. To disentangle the slow-changing aspects of each conditional from the fast-changing adaptations to each intervention, we parametrize the neural network into fast parameters and slow meta-parameters. We introduce a meta-learning objective that favours solutions \textit{robust} to frequent but sparse interventional distribution change, and which generalize well to previously unseen interventions. Optimizing this objective is shown experimentally to recover the structure of the causal graph. Finally, we find that when the learner is unaware of the intervention variable, it is able to infer that information, improving results further and focusing the parameter and meta-parameter updates where needed.

****

Adaptive Generation of Unrestricted Adversarial Inputs

Isaac Dunn,Hadrien Pouget,Tom Melham,Daniel Kroening

Neural networks are vulnerable to adversarially-constructed perturbations of their inputs. Most research so far has considered perturbations of a fixed magnitude under some l_p norm. Although studying these attacks is valuable, there has been increasing interest in the construction of—and robustness to—unrestricted attacks, which are not constrained to a small and rather artificial subset of all possible adversarial inputs. We introduce a novel algorithm for generating such unrestricted adversarial inputs which, unlike prior work, is adaptive: it is able to tune its attacks to the classifier being targeted. It also offers a 400–2,000× speedup over the existing state of the art. We demonstrate our approach by generating unrestricted adversarial inputs that fool classifiers robust to perturbation-based attacks. We also show that, by virtue of being adaptive and unrestricted, our attack is able to bypass adversarial training against it.

****

P-BN: Towards Effective Batch Normalization in the Path Space

Xufang Luo,Qi Meng,Wei Chen,Tie-Yan Liu

Neural networks with ReLU activation functions have demonstrated their success in many applications. Recently, researchers noticed a potential issue with the optimization of ReLU networks: the ReLU activation functions are positively scale-invariant (PSI), while the weights are not. This mismatch may lead to undesirable behaviors in the optimization process. Hence, some new algorithms that conduct optimizations directly in the path space (the path space is proven to be PSI) were developed, such as Stochastic Gradient Descent (SGD) in the path space, and it was shown that SGD in the path space is superior to that in the weight space. However, it is still unknown whether other deep learning techniques beyond SGD, such as batch normalization (BN), could also have their counterparts in the path space. In this paper, we conduct a formal study on the design of BN in the path space. According to our study, the key challenge is how to ensure the forward propagation in the path space, because BN is utilized during the forward process. To tackle such challenge, we propose a novel re-parameterization of ReLU networks, with which we replace each weight in the original neural network, with a new value calculated from one or several paths, while keeping the outputs of the network unchanged for any input. Then we show that BN in the path space, namely P-BN, is just a slightly modified conventional BN on the re-parameterized ReLU networks. Our experiments on two benchmark datasets, CIFAR and ImageNet, show that the proposed P-BN can signi■cantly outperform the conventional BN in the weight space.

****

Efficient Probabilistic Logic Reasoning with Graph Neural Networks

Yuyu Zhang,Xinshi Chen,Yuan Yang,Arun Ramamurthy,Bo Li,Yuan Qi,Le Song

Markov Logic Networks (MLNs), which elegantly combine logic rules and probabilistic graphical models, can be used to address many knowledge graph problems. However, inference in MLN is computationally intensive, making the industrial-scale application of MLN very difficult. In recent years, graph neural networks (GNNs)

have emerged as efficient and effective tools for large-scale graph problems. Nevertheless, GNNs do not explicitly incorporate prior logic rules into the models, and may require many labeled examples for a target task. In this paper, we explore the combination of MLNs and GNNs, and use graph neural networks for variational inference in MLN. We propose a GNN variant, named ExpressGNN, which strikes a nice balance between the representation power and the simplicity of the model. Our extensive experiments on several benchmark datasets demonstrate that ExpressGNN leads to effective and efficient probabilistic logic reasoning.
****
Low-dimensional statistical manifold embedding of directed graphs
Thorben Funke,Tian Guo,Alen Lancic,Nino Antulov-Fantulin
We propose a novel node embedding of directed graphs to statistical manifolds, which is based on a global minimization of pairwise relative entropy and graph geodesics in a non-linear way. Each node is encoded with a probability density function over a measurable space. Furthermore, we analyze the connection of the geometrical properties of such embedding and their efficient learning procedure. Extensive experiments show that our proposed embedding is better preserving the global geodesic information of graphs, as well as outperforming existing embedding models on directed graphs in a variety of evaluation metrics, in an unsupervised setting.
****
GATO: Gates Are Not the Only Option
Mark Goldstein*,Xintian Han*,Rajesh Ranganath
Recurrent Neural Networks (RNNs) facilitate prediction and generation of structured temporal data such as text and sound. However, training RNNs is hard. Vanishing gradients cause difficulties for learning long-range dependencies. Hidden states can explode for long sequences and send unbounded gradients to model parameters, even when hidden-to-hidden Jacobians are bounded. Models like the LSTM and GRU use gates to bound their hidden state, but most choices of gating functions lead to saturating gradients that contribute to, instead of alleviate, vanishing gradients. Moreover, performance of these models is not robust across random initializations. In this work, we specify desiderata for sequence models. We develop one model that satisfies them and that is capable of learning long-term dependencies, called GATO. GATO is constructed so that part of its hidden state does not have vanishing gradients, regardless of sequence length. We study GATO on copying and arithmetic tasks with long dependencies and on modeling intensive care unit and language data. Training GATO is more stable across random seeds and learning rates than GRUs and LSTMs. GATO solves these tasks using an order of magnitude fewer parameters.
****
Probabilistic View of Multi-agent Reinforcement Learning: A Unified Approach
Shubham Gupta,Ambedkar Dukkipati
Formulating the reinforcement learning (RL) problem in the framework of probabilistic inference not only offers a new perspective about RL, but also yields practical algorithms that are more robust and easier to train. While this connection between RL and probabilistic inference has been extensively studied in the single-agent setting, it has not yet been fully understood in the multi-agent setup. In this paper, we pose the problem of multi-agent reinforcement learning as the problem of performing inference in a particular graphical model. We model the environment, as seen by each of the agents, using separate but related Markov decision processes. We derive a practical off-policy maximum-entropy actor-critic algorithm that we call Multi-agent Soft Actor-Critic (MA-SAC) for performing approximate inference in the proposed model using variational inference. MA-SAC can be employed in both cooperative and competitive settings. Through experiments, we demonstrate that MA-SAC outperforms a strong baseline on several multi-agent scenarios. While MA-SAC is one resultant multi-agent RL algorithm that can be derived from the proposed probabilistic framework, our work provides a unified view of maximum-entropy algorithms in the multi-agent setting.
****
Neural Subgraph Isomorphism Counting

Xin Liu,Haojie Pan,Mutian He,Yangqiu Song,Xin Jiang
In this paper, we study a new graph learning problem: learning to count subgraph isomorphisms. Although the learning based approach is inexact, we are able to generalize to count large patterns and data graphs in polynomial time compared to the exponential time of the original NP-complete problem. Different from other traditional graph learning problems such as node classification and link prediction, subgraph isomorphism counting requires more global inference to oversee the whole graph. To tackle this problem, we propose a dynamic intermedium attention memory network (DIAMNet) which augments different representation learning architectures and iteratively attends pattern and target data graphs to memorize different subgraph isomorphisms for the global counting. We develop both small graphs (<= 1,024 subgraph isomorphisms in each) and large graphs (<= 4,096 subgraph isomorphisms in each) sets to evaluate different models. Experimental results show that learning based subgraph isomorphism counting can help reduce the time complexity with acceptable accuracy. Our DIAMNet can further improve existing representation learning models for this more global problem.
****

RIDE: Rewarding Impact-Driven Exploration for Procedurally-Generated Environments
Roberta Raileanu,Tim Rocktäschel
Exploration in sparse reward environments remains one of the key challenges of model-free reinforcement learning. Instead of solely relying on extrinsic rewards provided by the environment, many state-of-the-art methods use intrinsic rewards to encourage exploration. However, we show that existing methods fall short in procedurally-generated environments where an agent is unlikely to visit a state more than once. We propose a novel type of intrinsic reward which encourages the agent to take actions that lead to significant changes in its learned state representation. We evaluate our method on multiple challenging procedurally-generated tasks in MiniGrid, as well as on tasks with high-dimensional observations used in prior work. Our experiments demonstrate that this approach is more sample efficient than existing exploration methods, particularly for procedurally-generated MiniGrid environments. Furthermore, we analyze the learned behavior as well as the intrinsic reward received by our agent. In contrast to previous approaches, our intrinsic reward does not diminish during the course of training and it rewards the agent substantially more for interacting with objects that it can control.
****

Continual Learning with Delayed Feedback
THEIVENDIRAM PRANAVAN,TERENCE SIM
Most of the artificial neural networks are using the benefit of labeled datasets whereas in human brain, the learning is often unsupervised. The feedback or a label for a given input or a sensory stimuli is not often available instantly. After some time when brain gets the feedback, it updates its knowledge. That's how brain learns. Moreover, there is no training or testing phase. Human learns continually. This work proposes a model-agnostic continual learning framework which can be used with neural networks as well as decision trees to incorporate continual learning. Specifically, this work investigates how delayed feedback can be handled. In addition, a way to update the Machine Learning models with unlabeled data is proposed. Promising results are received from the experiments done on neural networks and decision trees.
****

Neural Non-additive Utility Aggregation
Markus Zopf
Neural architectures for set regression problems aim at learning representations such that good predictions can be made based on the learned representations. This strategy, however, ignores the fact that meaningful intermediate results might be helpful to perform well. We study two new architectures that explicitly model latent intermediate utilities and use non-additive utility aggregation to estimate the set utility based on the latent utilities. We evaluate the new architectures with visual and textual datasets, which have non-additive set utilities d

ue to redundancy and synergy effects. We find that the new architectures perform substantially better in this setup.
****
Bayesian Variational Autoencoders for Unsupervised Out-of-Distribution Detection
Erik Daxberger,José Miguel Hernández-Lobato
Despite their successes, deep neural networks still make unreliable predictions when faced with test data drawn from a distribution different to that of the training data, constituting a major problem for AI safety. While this motivated a recent surge in interest in developing methods to detect such out-of-distribution (OoD) inputs, a robust solution is still lacking. We propose a new probabilistic, unsupervised approach to this problem based on a Bayesian variational autoencoder model, which estimates a full posterior distribution over the decoder parameters using stochastic gradient Markov chain Monte Carlo, instead of fitting a point estimate. We describe how information-theoretic measures based on this posterior can then be used to detect OoD data both in input space as well as in the model's latent space. The effectiveness of our approach is empirically demonstrated.
****
``"Best-of-Many-Samples" Distribution Matching
Apratim Bhattacharyya,Mario Fritz,Bernt Schiele
Generative Adversarial Networks (GANs) can achieve state-of-the-art sample quality in generative modelling tasks but suffer from the mode collapse problem. Variational Autoencoders (VAE) on the other hand explicitly maximize a reconstruction-based data log-likelihood forcing it to cover all modes, but suffer from poorer sample quality. Recent works have proposed hybrid VAE-GAN frameworks which integrate a GAN-based synthetic likelihood to the VAE objective to address both the mode collapse and sample quality issues, with limited success. This is because the VAE objective forces a trade-off between the data log-likelihood and divergence to the latent prior. The synthetic likelihood ratio term also shows instability during training. We propose a novel objective with a ``"Best-of-Many-Samples" reconstruction cost and a stable direct estimate of the synthetic likelihood. This enables our hybrid VAE-GAN framework to achieve high data log-likelihood and low divergence to the latent prior at the same time and shows significant improvement over both hybrid VAE-GANS and plain GANs in mode coverage and quality.
****
SPACE: Unsupervised Object-Oriented Scene Representation via Spatial Attention and Decomposition
Zhixuan Lin,Yi-Fu Wu,Skand Vishwanath Peri,Weihao Sun,Gautam Singh,Fei Deng,Jindong Jiang,Sungjin Ahn
The ability to decompose complex multi-object scenes into meaningful abstractions like objects is fundamental to achieve higher-level cognition. Previous approaches for unsupervised object-oriented scene representation learning are either based on spatial-attention or scene-mixture approaches and limited in scalability which is a main obstacle towards modeling real-world scenes. In this paper, we propose a generative latent variable model, called SPACE, that provides a uni■ed probabilistic modeling framework that combines the best of spatial-attention and scene-mixture approaches. SPACE can explicitly provide factorized object representations for foreground objects while also decomposing background segments of complex morphology. Previous models are good at either of these, but not both. SPACE also resolves the scalability problems of previous methods by incorporating parallel spatial-attention and thus is applicable to scenes with a large number of objects without performance degradations. We show through experiments on Atari and 3D-Rooms that SPACE achieves the above properties consistently in comparison to SPAIR, IODINE, and GENESIS. Results of our experiments can be found on our project website: https://sites.google.com/view/space-project-page
****
Efficient generation of structured objects with Constrained Adversarial Networks
Jacopo Gobbi,Luca Di Liello,Pierfrancesco Ardino,Paolo Morettin,Stefano Teso,Andrea Passerini
Despite their success, generative adversarial networks (GANs) cannot easily gene

rate structured objects like molecules or game maps. The issue is that such objects must satisfy structural requirements (e.g., molecules must be chemically valid, game maps must guarantee reachability of the end goal) that are difficult to capture with examples alone. As a remedy, we propose constrained adversarial networks (CANs), which embed the constraints into the model during training by penalizing the generator whenever it outputs invalid structures. As in unconstrained GANs, new objects can be sampled straightforwardly from the generator, but in addition they satisfy the constraints with high probability. Our approach handles arbitrary logical constraints and leverages knowledge compilation techniques to efficiently evaluate the expected disagreement between the model and the constraints. This setup is further extended to hybrid logical-neural constraints for capturing complex requirements like graph reachability. An extensive empirical analysis on constrained images, molecules, and video game levels shows that CANs efficiently generate valid structures that are both high-quality and novel.

****

Deep Variational Semi-Supervised Novelty Detection

Tal Daniel,Thanard Kurutach,Aviv Tamar

In anomaly detection (AD), one seeks to identify whether a test sample is abnormal, given a data set of normal samples. A recent and promising approach to AD relies on deep generative models, such as variational autoencoders (VAEs),for unsupervised learning of the normal data distribution. In semi-supervised AD (SSAD), the data also includes a small sample of labeled anomalies. In this work,we propose two variational methods for training VAEs for SSAD. The intuitive idea in both methods is to train the encoder to ‘separate’ between latent vectors for normal and outlier data. We show that this idea can be derived from principled probabilistic formulations of the problem, and propose simple and effective algorithms. Our methods can be applied to various data types, as we demonstrate on SSAD datasets ranging from natural images to astronomy and medicine, and can be combined with any VAE model architecture. When comparing to state-of-the-art SSAD methods that are not specific to particular data types, we obtain marked improvement in outlier detection.

****

Cross-Lingual Ability of Multilingual BERT: An Empirical Study

Karthikeyan K,Zihan Wang,Stephen Mayhew,Dan Roth

Recent work has exhibited the surprising cross-lingual abilities of multilingual BERT (M-BERT) -- surprising since it is trained without any cross-lingual objective and with no aligned data. In this work, we provide a comprehensive study of the contribution of different components in M-BERT to its cross-lingual ability. We study the impact of linguistic properties of the languages, the architecture of the model, and the learning objectives. The experimental study is done in the context of three typologically different languages -- Spanish, Hindi, and Russian -- and using two conceptually different NLP tasks, textual entailment and named entity recognition. Among our key conclusions is the fact that the lexical overlap between languages plays a negligible role in the cross-lingual success, while the depth of the network is an integral part of it. All our models and implementations can be found on our project page: http://cogcomp.org/page/publication_view/900.

****

Towards Finding Longer Proofs

Zsolt Zombori,Adrián Csiszárik,Henryk Michalewski,Cezary Kaliszyk,Josef Urban

We present a reinforcement learning (RL) based guidance system for automated theorem proving geared towards Finding Longer Proofs (FLoP). FLoP focuses on generalizing from short proofs to longer ones of similar structure. To achieve that, FLoP uses state-of-the-art RL approaches that were previously not applied in theorem proving. In particular, we show that curriculum learning significantly outperforms previous learning-based proof guidance on a synthetic dataset of increasingly difficult arithmetic problems.

****

Probing Emergent Semantics in Predictive Agents via Question Answering

Abhishek Das,Federico Carnevale,Hamza Merzic,Laura Rimell,Rosalia Schneider,Alden Hung,Josh Abramson,Arun Ahuja,Stephen Clark,Greg Wayne,Felix Hill
Recent work has demonstrated how predictive modeling can endow agents with rich knowledge of their surroundings, improving their ability to act in complex environments. We propose question-answering as a general paradigm to decode and understand the representations that such agents develop, applying our method to two recent approaches to predictive modeling – action-conditional CPC (Guo et al., 2018) and SimCore (Gregor et al., 2019). After training agents with these predictive objectives in a visually-rich, 3D environment with an assortment of objects, colors, shapes, and spatial configurations, we probe their internal state representations with a host of synthetic (English) questions, without backpropagating gradients from the question-answering decoder into the agent. The performance of different agents when probed in this way reveals that they learn to encode detailed, and seemingly compositional, information about objects, properties and spatial relations from their physical environment. Our approach is intuitive, i.e. humans can easily interpret the responses of the model as opposed to inspecting continuous vectors, and model-agnostic, i.e. applicable to any modeling approach. By revealing the implicit knowledge of objects, quantities, properties and relations acquired by agents as they learn, question-conditional agent probing can stimulate the design and development of stronger predictive learning objectives.
****
Hierarchical Graph Matching Networks for Deep Graph Similarity Learning
Xiang Ling,Lingfei Wu,Saizhuo Wang,Tengfei Ma,Fangli Xu,Chunming Wu,Shouling Ji
While the celebrated graph neural networks yields effective representations for individual nodes of a graph, there has been relatively less success in extending to deep graph similarity learning.
Recent work has considered either global-level graph-graph interactions or low-level node-node interactions, ignoring the rich cross-level interactions between parts of a graph and a whole graph.
In this paper, we propose a Hierarchical Graph Matching Network (HGMN) for computing the graph similarity between any pair of graph-structured objects. Our model jointly learns graph representations and a graph matching metric function for computing graph similarity in an end-to-end fashion. The proposed HGMN model consists of a multi-perspective node-graph matching network for effectively learning cross-level interactions between parts of a graph and a whole graph, and a siamese graph neural network for learning global-level interactions between two graphs. Our comprehensive experiments demonstrate that our proposed HGMN consistently outperforms state-of-the-art graph matching networks baselines for both classification and regression tasks.
****
A Simple Approach to the Noisy Label Problem Through the Gambler's Loss
Liu Ziyin,Ru Wang,Paul Pu Liang,Ruslan Salakhutdinov,Louis-Philippe Morency,Masahito Ueda
Learning in the presence of label noise is a challenging yet important task. It is crucial to design models that are robust to noisy labels. In this paper, we discover that a new class of loss functions called the gambler's loss provides strong robustness to label noise across various levels of corruption. Training with this modified loss function reduces memorization of data points with noisy labels and is a simple yet effective method to improve robustness and generalization. Moreover, using this loss function allows us to derive an analytical early stopping criterion that accurately estimates when memorization of noisy labels begins to occur. Our overall approach achieves strong results and outperforming existing baselines.
****
On the Reflection of Sensitivity in the Generalization Error
Mahsa Forouzesh,Farnood Salehi,Patrick Thiran
Even though recent works have brought some insight into the performance improvement of techniques used in state-of-the-art deep-learning models, more work is needed to understand the generalization properties of over-parameterized deep neural networks. We shed light on this matter by linking the loss function to the ou

tput’s sensitivity to its input. We find a rather strong empirical relation betw
een the output sensitivity and the variance in the bias-variance decomposition o
f the loss function, which hints on using sensitivity as a metric for comparing
generalization performance of networks, without requiring labeled data. We find
that sensitivity is decreased by applying popular methods which improve the gene
ralization performance of the model, such as (1) using a deep network rather tha
n a wide one, (2) adding convolutional layers to baseline classifiers instead of
 adding fully connected layers, (3) using batch normalization, dropout and max-p
ooling, and (4) applying parameter initialization techniques.
****

Redundancy-Free Computation Graphs for Graph Neural Networks
Zhihao Jia,Sina Lin,Rex Ying,Jiaxuan You,Jure Leskovec,Alex Aiken.
Graph Neural Networks (GNNs) are based on repeated aggregations of information a
cross nodes’ neighbors in a graph. However, because common neighbors are shared
 between different nodes, this leads to repeated and inefficient computations.We
 propose Hierarchically Aggregated computation Graphs (HAGs), a new GNN graph re
presentation that explicitly avoids redundancy by managing intermediate aggregat
ion results hierarchically, and eliminating repeated computations and unn
ecessary data transfers in GNN training and inference. We introduce an accurate
cost function to quantitatively evaluate the runtime performance of different HA
Gsand use a novel search algorithm to find optimized HAGs. Experiments show tha
t the HAG representation significantly outperforms the standard GNN graph repres
entation by increasing the end-to-end training throughput by up to 2.8x and redu
cing the aggregations and data transfers in GNN training by up to 6.3x and 5.6x.
 Meanwhile, HAGs improve runtime performance by preserving GNNcomputation, and m
aintain the original model accuracy for arbitrary GNNs.
****

Toward Understanding The Effect of Loss Function on The Performance of Knowledge
 Graph Embedding
Mojtaba Nayyeri,Chengjin Xu,Yadollah Yaghoobzadeh,Hamed Shariat Yazdi,Jens Lehma
nn
Knowledge graphs (KGs) represent world's facts in structured forms. KG completi
on exploits the existing facts in a KG to discover new ones. Translation-based e
mbedding model (TransE) is a prominent formulation to do KG completion.
Despite the efficiency of TransE in memory and time, it suffers from several lim
itations in encoding relation patterns such as symmetric, reflexive etc. To res
olve this problem, most of the attempts have circled around the revision of the
score function of TransE i.e., proposing a more complicated score function such
as Trans(A, D, G, H, R, etc) to mitigate the limitations. In this paper, we tac
kle this problem from a different perspective. We show that existing theories co
rresponding to the limitations of TransE are inaccurate because they ignore the
effect of loss function. Accordingly, we pose theoretical investigations of the
main limitations of TransE in the light of loss function. To the best of our kno
wledge, this has not been investigated so far comprehensively. We show that by a
 proper selection of the loss function for training the TransE model, the main l
imitations of the model are mitigated. This is explained by setting upper-bound
for the scores of positive samples, showing the region of truth (i.e., the regio
n that a triple is considered positive by the model).
Our theoretical proofs with experimental results fill the gap between the capabi
lity of translation-based class of embedding models and the loss function. The t
heories emphasis the importance of the selection of the loss functions for train
ing the models. Our experimental evaluations on different loss functions used fo
r training the models justify our theoretical proofs and confirm the importance
of the loss functions on the performance.

****

Reducing Transformer Depth on Demand with Structured Dropout
Angela Fan,Edouard Grave,Armand Joulin
Overparametrized transformer networks have obtained state of the art results in

various natural language processing tasks, such as machine translation, language modeling, and question answering. These models contain hundreds of millions of parameters, necessitating a large amount of computation■and making them prone to overfitting. In this work, we explore LayerDrop, a form of structured dropout, which has a regularization effect during training and allows for efficient pruning at inference time. In particular, we show that it is possible to select sub-networks of any depth from one large network without having to finetune them and with limited impact on performance. We demonstrate the effectiveness of our■approach by improving the state of the art on machine translation, language modeling, summarization, question answering, and language understanding benchmarks. Moreover, we show that our approach leads to small BERT-like models of higher quality than when training from scratch or using distillation.
****
Semi-Supervised Learning with Normalizing Flows
Pavel Izmailov,Polina Kirichenko,Marc Finzi,Andrew Wilson
We propose Flow Gaussian Mixture Model (FlowGMM), a general-purpose method for semi-supervised learning based on a simple and principled probabilistic framework. We approximate the joint distribution of the labeled and unlabeled data with a flexible mixture model implemented as a Gaussian mixture transformed by a normalizing flow. We train the model by maximizing the exact joint likelihood of the labeled and unlabeled data. We evaluate FlowGMM on a wide range of semi-supervised classification problems across different data types: AG-News and Yahoo Answers text data, MNIST, SVHN and CIFAR-10 image classification problems as well as tabular UCI datasets. FlowGMM achieves promising results on image classification problems and outperforms the competing methods on other types of data. FlowGMM learns an interpretable latent repesentation space and allows hyper-parameter free feature visualization at real time rates. Finally, we show that FlowGMM can be calibrated to produce meaningful uncertainty estimates for its predictions.
****
Neural Communication Systems with Bandwidth-limited Channel
Karen Ullrich,Fabio Viola,Danilo J. Rezende
Reliably transmitting messages despite information loss due to a noisy channel is a core problem of information theory. One of the most important aspects of real world communication is that it may happen at varying levels of information transfer. The bandwidth-limited channel models this phenomenon. In this study we consider learning joint coding with the bandwidth-limited channel. Although, classical results suggest that it is asymptotically optimal to separate the sub-tasks of compression (source coding) and error correction (channel coding), it is well known that for finite block-length problems, and when there are restrictions to the computational complexity of coding, this optimality may not be achieved. Thus, we empirically compare the performance of joint and separate systems, and conclude that joint systems outperform their separate counterparts when coding is performed by flexible learnable function approximators such as neural networks. Specifically, we cast the joint communication problem as a variational learning problem. To facilitate this, we introduce a differentiable and computationally efficient version of this channel. We show that our design compensates for the loss of information by two mechanisms: (i) missing information is modelled by a prior model incorporated in the channel model, and (ii) sampling from the joint model is improved by auxiliary latent variables in the decoder. Experimental results justify the validity of our design decisions through improved distortion and FID scores.
****
Reducing Computation in Recurrent Networks by Selectively Updating State Neurons
Thomas Hartvigsen,Cansu Sen,Xiangnan Kong,Elke Rundensteiner
Recurrent Neural Networks (RNN) are the state-of-the-art approach to sequential learning. However, standard RNNs use the same amount of computation at each time step, regardless of the input data. As a result, even for high-dimensional hidden states, all dimensions are updated at each timestep regardless of the recurrent memory cell. Reducing this rigid assumption could allow for models with large hidden states to perform inference more quickly. Intuitively, not all hidden sta

te dimensions need to be recomputed from scratch at each timestep. Thus, recent methods have begun studying this problem by imposing mainly a priori-determined patterns for updating the state. In contrast, we now design a fully-learned approach, SA-RNN, that augments any RNN by predicting discrete update patterns at the fine granularity of independent hidden state dimensions through the parameterization of a distribution of update-likelihoods driven entirely by the input data. We achieve this without imposing assumptions on the structure of the update pattern. Better yet, our method adapts the update patterns online, allowing different dimensions to be updated conditional to the input. To learn which to update, the model solves a multi-objective optimization problem, maximizing accuracy while minimizing the number of updates based on a unified control. Using publicly-available datasets we demonstrate that our method consistently achieves higher accuracy with fewer updates compared to state-of-the-art alternatives. Additionally, our method can be directly applied to a wide variety of models containing RNN architectures.
****
A Novel Analysis Framework of Lower Complexity Bounds for Finite-Sum Optimization
Guangzeng Xie,Luo Luo,Zhihua Zhang
This paper studies the lower bound complexity for the optimization problem whose objective function is the average of n individual smooth convex functions. We consider the algorithm which gets access to gradient and proximal oracle for each individual component.
For the strongly-convex case, we prove such an algorithm can not reach an $\eps$-suboptimal point in fewer than $\Omega((n+\sqrt{\kappa n})\log(1/\eps))$ iterations, where κ is the condition number of the objective function. This lower bound is tighter than previous results and perfectly matches the upper bound of the existing proximal incremental first-order oracle algorithm Point-SAGA.
We develop a novel construction to show the above result, which partitions the tridiagonal matrix of classical examples into n groups to make the problem difficult enough to stochastic algorithms.
This construction is friendly to the analysis of proximal oracle and also could be used in general convex and average smooth cases naturally.
****
Neural Outlier Rejection for Self-Supervised Keypoint Learning
Jiexiong Tang,Hanme Kim,Vitor Guizilini,Sudeep Pillai,Rares Ambrus
Identifying salient points in images is a crucial component for visual odometry, Structure-from-Motion or SLAM algorithms. Recently, several learned keypoint methods have demonstrated compelling performance on challenging benchmarks. However, generating consistent and accurate training data for interest-point detection in natural images still remains challenging, especially for human annotators.
We introduce IO-Net (i.e. InlierOutlierNet), a novel proxy task for the self-supervision of keypoint detection, description and matching. By making the sampling of inlier-outlier sets from point-pair correspondences fully differentiable within the keypoint learning framework, we show that are able to simultaneously self-supervise keypoint description and improve keypoint matching. Second, we introduce KeyPointNet, a keypoint-network architecture that is especially amenable to robust keypoint detection and description. We design the network to allow local keypoint aggregation to avoid artifacts due to spatial discretizations commonly used for this task, and we improve fine-grained keypoint descriptor performance by taking advantage of efficient sub-pixel convolutions to upsample the descriptor feature-maps to a higher operating resolution. Through extensive experiments and ablative analysis, we show that the proposed self-supervised keypoint learning method greatly improves the quality of feature matching and homography estimation on challenging benchmarks over the state-of-the-art.
****
B-Spline CNNs on Lie groups
Erik J Bekkers
Group convolutional neural networks (G-CNNs) can be used to improve classical CNNs by equipping them with the geometric structure of groups. Central in the succ

ess of G-CNNs is the lifting of feature maps to higher dimensional disentangled representations, in which data characteristics are effectively learned, geometric data-augmentations are made obsolete, and predictable behavior under geometric transformations (equivariance) is guaranteed via group theory. Currently, however, the practical implementations of G-CNNs are limited to either discrete groups (that leave the grid intact) or continuous compact groups such as rotations (that enable the use of Fourier theory). In this paper we lift these limitations and propose a modular framework for the design and implementation of G-CNNs for arbitrary Lie groups. In our approach the differential structure of Lie groups is used to expand convolution kernels in a generic basis of B-splines that is defined on the Lie algebra. This leads to a flexible framework that enables localized, atrous, and deformable convolutions in G-CNNs by means of respectively localized, sparse and non-uniform B-spline expansions. The impact and potential of our approach is studied on two benchmark datasets: cancer detection in histopathology slides (PCam dataset) in which rotation equivariance plays a key role and facial landmark localization (CelebA dataset) in which scale equivariance is important. In both cases, G-CNN architectures outperform their classical 2D counterparts and the added value of atrous and localized group convolutions is studied in detail.

****

Quantifying Point-Prediction Uncertainty in Neural Networks via Residual Estimation with an I/O Kernel

Xin Qiu,Elliot Meyerson,Risto Miikkulainen

Neural Networks (NNs) have been extensively used for a wide spectrum of real-world regression tasks, where the goal is to predict a numerical outcome such as revenue, effectiveness, or a quantitative result. In many such tasks, the point prediction is not enough: the uncertainty (i.e. risk or confidence) of that prediction must also be estimated. Standard NNs, which are most often used in such tasks, do not provide uncertainty information. Existing approaches address this issue by combining Bayesian models with NNs, but these models are hard to implement, more expensive to train, and usually do not predict as accurately as standard NNs. In this paper, a new framework (RIO) is developed that makes it possible to estimate uncertainty in any pretrained standard NN. The behavior of the NN is captured by modeling its prediction residuals with a Gaussian Process, whose kernel includes both the NN's input and its output. The framework is justified theoretically and evaluated in twelve real-world datasets, where it is found to (1) provide reliable estimates of uncertainty, (2) reduce the error of the point predictions, and (3) scale well to large datasets. Given that RIO can be applied to any standard NN without modifications to model architecture or training pipeline, it provides an important ingredient for building real-world NN applications.

****

EMPIR: Ensembles of Mixed Precision Deep Networks for Increased Robustness Against Adversarial Attacks

Sanchari Sen,Balaraman Ravindran,Anand Raghunathan

Ensuring robustness of Deep Neural Networks (DNNs) is crucial to their adoption in safety-critical applications such as self-driving cars, drones, and healthcare. Notably, DNNs are vulnerable to adversarial attacks in which small input perturbations can produce catastrophic misclassifications. In this work, we propose EMPIR, ensembles of quantized DNN models with different numerical precisions, as a new approach to increase robustness against adversarial attacks. EMPIR is based on the observation that quantized neural networks often demonstrate much higher robustness to adversarial attacks than full precision networks, but at the cost of a substantial loss in accuracy on the original (unperturbed) inputs. EMPIR overcomes this limitation to achieve the ``best of both worlds", i.e., the higher unperturbed accuracies of the full precision models combined with the higher robustness of the low precision models, by composing them in an ensemble. Further, as low precision DNN models have significantly lower computational and storage requirements than full precision models, EMPIR models only incur modest compute and memory overheads compared to a single full-precision model (<25% in our evaluations). We evaluate EMPIR across a suite of DNNs for 3 different image recog

nition tasks (MNIST, CIFAR-10 and ImageNet) and under 4 different adversarial attacks. Our results indicate that EMPIR boosts the average adversarial accuracies by 42.6%, 15.2% and 10.5% for the DNN models trained on the MNIST, CIFAR-10 and ImageNet datasets respectively, when compared to single full-precision models, without sacrificing accuracy on the unperturbed inputs.
****

Learning To Explore Using Active Neural SLAM
Devendra Singh Chaplot,Dhiraj Gandhi,Saurabh Gupta,Abhinav Gupta,Ruslan Salakhutdinov
This work presents a modular and hierarchical approach to learn policies for exploring 3D environments, called `Active Neural SLAM'. Our approach leverages the strengths of both classical and learning-based methods, by using analytical path planners with learned SLAM module, and global and local policies. The use of learning provides flexibility with respect to input modalities (in the SLAM module), leverages structural regularities of the world (in global policies), and provides robustness to errors in state estimation (in local policies). Such use of learning within each module retains its benefits, while at the same time, hierarchical decomposition and modular training allow us to sidestep the high sample complexities associated with training end-to-end policies. Our experiments in visually and physically realistic simulated 3D environments demonstrate the effectiveness of our approach over past learning and geometry-based approaches. The proposed model can also be easily transferred to the PointGoal task and was the winning entry of the CVPR 2019 Habitat PointGoal Navigation Challenge.
****

Adversarial Robustness Against the Union of Multiple Perturbation Models
Pratyush Maini,Eric Wong,Zico Kolter
Owing to the susceptibility of deep learning systems to adversarial attacks, there has been a great deal of work in developing (both empirically and certifiably) robust classifiers, but the vast majority has defended against single types of attacks. Recent work has looked at defending against multiple attacks, specifically on the MNIST dataset, yet this approach used a relatively complex architecture, claiming that standard adversarial training can not apply because it "overfits" to a particular norm. In this work, we show that it is indeed possible to adversarially train a robust model against a union of norm-bounded attacks, by using a natural generalization of the standard PGD-based procedure for adversarial training to multiple threat models. With this approach, we are able to train standard architectures which are robust against l_inf, l_2, and l_1 attacks, outperforming past approaches on the MNIST dataset and providing the first CIFAR10 network trained to be simultaneously robust against (l_inf, l_2, l_1) threat models, which achieves adversarial accuracy rates of (47.6%, 64.3%, 53.4%) for (l_inf, l_2, l_1) perturbations with epsilon radius = (0.03,0.5,12).
****

Understanding and Improving Information Transfer in Multi-Task Learning
Sen Wu,Hongyang R. Zhang,Christopher Ré
We investigate multi-task learning approaches that use a shared feature representation for all tasks. To better understand the transfer of task information, we study an architecture with a shared module for all tasks and a separate output module for each task. We study the theory of this setting on linear and ReLU-activated models. Our key observation is that whether or not tasks' data are well-aligned can significantly affect the performance of multi-task learning. We show that misalignment between task data can cause negative transfer (or hurt performance) and provide sufficient conditions for positive transfer. Inspired by the theoretical insights, we show that aligning tasks' embedding layers leads to performance gains for multi-task training and transfer learning on the GLUE benchmark and sentiment analysis tasks; for example, we obtained a 2.35% GLUE score average improvement on 5 GLUE tasks over BERT LARGE using our alignment method. We also design an SVD-based task re-weighting scheme and show that it improves the robustness of multi-task training on a multi-label image dataset.
****

Hyperparameter Tuning and Implicit Regularization in Minibatch SGD

Samuel L Smith,Erich Elsen,Soham De
This paper makes two contributions towards understanding how the hyperparameters of stochastic gradient descent affect the final training loss and test accuracy of neural networks. First, we argue that stochastic gradient descent exhibits two regimes with different behaviours; a noise dominated regime which typically arises for small or moderate batch sizes, and a curvature dominated regime which typically arises when the batch size is large. In the noise dominated regime, the optimal learning rate increases as the batch size rises, and the training loss and test accuracy are independent of batch size under a constant epoch budget. In the curvature dominated regime, the optimal learning rate is independent of batch size, and the training loss and test accuracy degrade as the batch size rises. We support these claims with experiments on a range of architectures including ResNets, LSTMs and autoencoders. We always perform a grid search over learning rates at all batch sizes. Second, we demonstrate that small or moderately large batch sizes continue to outperform very large batches on the test set, even when both models are trained for the same number of steps and reach similar training losses. Furthermore, when training Wide-ResNets on CIFAR-10 with a constant batch size of 64, the optimal learning rate to maximize the test accuracy only decays by a factor of 2 when the epoch budget is increased by a factor of 128, while the optimal learning rate to minimize the training loss decays by a factor of 16. These results confirm that the noise in stochastic gradients can introduce beneficial implicit regularization.
****

Searching for Stage-wise Neural Graphs In the Limit
Xin Zhou,Dejing Dou,Boyang Li
Search space is a key consideration for neural architecture search. Recently, Xie et al. (2019a) found that randomly generated networks from the same distribution perform similarly, which suggest we should search for random graph distributions instead of graphs. We propose graphon as a new search space. A graphon is the limit of Cauchy sequence of graphs and a scale-free probabilistic distribution, from which graphs of different number of vertices can be drawn. This property enables us to perform NAS using fast, low-capacity models and scale the found models up when necessary. We develop an algorithm for NAS in the space of graphons and empirically demonstrate that it can find stage-wise graphs that outperform DenseNet and other baselines on ImageNet.
****

Restricting the Flow: Information Bottlenecks for Attribution
Karl Schulz,Leon Sixt,Federico Tombari,Tim Landgraf
Attribution methods provide insights into the decision-making of machine learning models like artificial neural networks. For a given input sample, they assign a relevance score to each individual input variable, such as the pixels of an image. In this work, we adopt the information bottleneck concept for attribution. By adding noise to intermediate feature maps, we restrict the flow of information and can quantify (in bits) how much information image regions provide. We compare our method against ten baselines using three different metrics on VGG-16 and ResNet-50, and find that our methods outperform all baselines in five out of six settings. The method’s information-theoretic foundation provides an absolute frame of reference for attribution values (bits) and a guarantee that regions scored close to zero are not necessary for the network's decision.
****

Stein Bridging: Enabling Mutual Reinforcement between Explicit and Implicit Generative Models
Qitian Wu,Rui Gao,Hongyuan Zha
Deep generative models are generally categorized into explicit models and implicit models. The former assumes an explicit density form whose normalizing constant is often unknown; while the latter, including generative adversarial networks (GANs), generates samples using a push-forward mapping. In spite of substantial recent advances demonstrating the power of the two classes of generative models in many applications, both of them, when used alone, suffer from respective limitations and drawbacks. To mitigate these issues, we propose Stein Bridging, a no

vel joint training framework that connects an explicit density estimator and an implicit sample generator with Stein discrepancy. We show that the Stein Bridge induces new regularization schemes for both explicit and implicit models. Conver gence analysis and extensive experiments demonstrate that the Stein Bridging i) improves the stability and sample quality of the GAN training, and ii) facilitat es the density estimator to seek more modes in data and alleviate the mode-colla pse issue. Additionally, we discuss several applications of Stein Bridging and u seful tricks in practical implementation used in our experiments.
****
Step Size Optimization
Gyoung S. Na,Dongmin Hyeon,Hwanjo Yu
This paper proposes a new approach for step size adaptation in gradient methods. The proposed method called step size optimization (SSO) formulates the step siz e adaptation as an optimization problem which minimizes the loss function with r espect to the step size for the given model parameters and gradients. Then, the step size is optimized based on alternating direction method of multipliers (ADM M). SSO does not require the second-order information or any probabilistic model s for adapting the step size, so it is efficient and easy to implement. Furtherm ore, we also introduce stochastic SSO for stochastic learning environments. In t he experiments, we integrated SSO to vanilla SGD and Adam, and they outperformed state-of-the-art adaptive gradient methods including RMSProp, Adam, L4-Adam, an d AdaBound on extensive benchmark datasets.
****
Equilibrium Propagation with Continual Weight Updates
Maxence Ernoult,Julie Grollier,Damien Querlioz,Yoshua Bengio,Benjamin Scellier
Equilibrium Propagation (EP) is a learning algorithm that bridges Machine Learni ng and Neuroscience, by computing gradients closely matching those of Backpropag ation Through Time (BPTT), but with a learning rule local in space.
Given an input x and associated target y, EP proceeds in two phases: in the firs t phase neurons evolve freely towards a first steady state; in the second phase output neurons are nudged towards y until they reach a second steady state.
However, in existing implementations of EP, the learning rule is not local in ti me:
the weight update is performed after the dynamics of the second phase have conve rged and requires information of the first phase that is no longer available phy sically.
This is a major impediment to the biological plausibility of EP and its efficien t hardware implementation.
In this work, we propose a version of EP named Continual Equilibrium Propagation (C-EP) where neuron and synapse dynamics occur simultaneously throughout the se cond phase, so that the weight update becomes local in time. We prove theoretica lly that, provided the learning rates are sufficiently small, at each time step of the second phase the dynamics of neurons and synapses follow the gradients of the loss given by BPTT (Theorem 1).
We demonstrate training with C-EP on MNIST and generalize C-EP to neural network s where neurons are connected by asymmetric connections. We show through experim ents that the more the network updates follows the gradients of BPTT, the best i t performs in terms of training. These results bring EP a step closer to biology while maintaining its intimate link with backpropagation.
****
Global Adversarial Robustness Guarantees for Neural Networks
Luca Laurenti,Andrea Patane,Matthew Wicker,Luca Bortolussi,Luca Cardelli,Marta K wiatkowska
We investigate global adversarial robustness guarantees for machine learning mod els. Specifically, given a trained model we consider the problem of computing t he probability that its prediction at any point sampled from the (unknown) input distribution is susceptible to adversarial attacks. Assuming continuity of the model, we prove measurability for a selection of local robustness properties us ed in the literature. We then show how concentration inequalities can be employe d to compute global robustness with estimation error upper-bounded by ϵ

, for any $\epsilon > 0$ selected a priori. We utilise the methods to provide st
atistically sound analysis of the robustness/accuracy trade-off for a variety of
 neural networks architectures and training methods on MNIST, Fashion-MNIST and
CIFAR. We empirically observe that robustness and accuracy tend to be negatively
 correlated for networks trained via stochastic gradient descent and with iterat
ive pruning techniques, while a positive trend is observed between them in Bayes
ian settings.
****

A Stochastic Derivative Free Optimization Method with Momentum
Eduard Gorbunov,Adel Bibi,Ozan Sener,El Houcine Bergou,Peter Richtarik
We consider the problem of unconstrained minimization of a smooth objective
function in $\mathbb{R}^d$ in setting where only function evaluations are possib
le. We propose and analyze stochastic zeroth-order method with heavy ball moment
um. In particular, we propose, SMTP, a momentum version of the stochastic three-
point method (STP) Bergou et al. (2019). We show new complexity results for non-
convex, convex and strongly convex functions. We test our method on a collection
 of learning to continuous control tasks on several MuJoCo Todorov et al. (2012)
 environments with varying difficulty and compare against STP, other state-of-th
e-art derivative-free optimization algorithms and against policy gradient method
s. SMTP significantly outperforms STP and all other methods that we considered i
n our numerical experiments. Our second contribution is SMTP with importance sam
pling which we call SMTP_IS. We provide convergence analysis of this method for
non-convex, convex and strongly convex objectives.
****

Coresets for Accelerating Incremental Gradient Methods
Baharan Mirzasoleiman,Jeff Bilmes,Jure Leskovec
Many machine learning problems reduce to the problem of minimizing an expected r
isk. Incremental gradient (IG) methods, such as stochastic gradient descent and
its variants, have been successfully used to train the largest of machine learni
ng models. IG methods, however, are in general slow to converge and sensitive t
o stepsize choices. Therefore, much work has focused on speeding them up by redu
cing the variance of the estimated gradient or choosing better stepsizes. An alt
ernative strategy would be to select a carefully chosen subset of training data,
 train only on that subset, and hence speed up optimization. However, it remains
 an open question how to achieve this, both theoretically as well as practically
, while not compromising on the quality of the final model. Here we develop CRA
IG, a method for selecting a weighted subset (or coreset) of training data in or
der to speed up IG methods. We prove that by greedily selecting a subset S of tr
aining data that minimizes the upper-bound on the estimation error of the full g
radient, running IG on this subset will converge to the (near)optimal solution i
n the same number of epochs as running IG on the full data. But because at each
epoch the gradients are computed only on the subset S, we obtain a speedup that
is inversely proportional to the size of S. Our subset selection algorithm is fu
lly general and can be applied to most IG methods. We further demonstrate practi
cal effectiveness of our algorithm, CRAIG, through an extensive set of experimen
ts on several applications, including logistic regression and deep neural networ
ks. Experiments show that CRAIG, while achieving practically the same loss, spee
ds up IG methods by up to 10x for convex and 3x for non-convex (deep learning) p
roblems.
****

A Greedy Approach to Max-Sliced Wasserstein GANs
András Horváth
Generative Adversarial Networks have made data generation possible in various us
e cases, but in case of complex, high-dimensional distributions it can be diffic
ult to train them, because of convergence problems and the appearance of mode co
llapse.
Sliced Wasserstein GANs and especially the application of the Max-Sliced Wassers
tein distance made it possible to approximate Wasserstein distance during traini
ng in an efficient and stable way and helped ease convergence problems of these
architectures.

This method transforms sample assignment and distance calculation into sorting t
he one-dimensional projection of the samples, which results a sufficient approxi
mation of the high-dimensional Wasserstein distance.

In this paper we will demonstrate that the approximation of the Wasserstein dist
ance by sorting the samples is not always the optimal approach and the greedy as
signment of the real and fake samples can result faster convergence and better a
pproximation of the original distribution.
****
Off-Policy Actor-Critic with Shared Experience Replay
Simon Schmitt,Matteo Hessel,Karen Simonyan
We investigate the combination of actor-critic reinforcement learning algorithms
 with uniform large-scale experience replay and propose solutions for two challe
nges: (a) efficient actor-critic learning with experience replay (b) stability o
f very off-policy learning. We employ those insights to accelerate hyper-paramet
er sweeps in which all participating agents run concurrently and share their exp
erience via a common replay module.

To this end we analyze the bias-variance tradeoffs in V-trace, a form of importa
nce sampling for actor-critic methods. Based on our analysis, we then argue for
mixing experience sampled from replay with on-policy experience, and propose a n
ew trust region scheme that scales effectively to data distributions where V-tra
ce becomes unstable.

We provide extensive empirical validation of the proposed solution. We further s
how the benefits of this setup by demonstrating state-of-the-art data efficiency
 on Atari among agents trained up until 200M environment frames.
****
Intrinsically Motivated Discovery of Diverse Patterns in Self-Organizing Systems
Chris Reinke,Mayalen Etcheverry,Pierre-Yves Oudeyer
In many complex dynamical systems, artificial or natural, one can observe self-o
rganization of patterns emerging from local rules. Cellular automata, like the G
ame of Life (GOL), have been widely used as abstract models enabling the study o
f various aspects of self-organization and morphogenesis, such as the emergence
of spatially localized patterns. However, findings of self-organized patterns in
 such models have so far relied on manual tuning of parameters and initial state
s, and on the human eye to identify interesting patterns. In this paper, we form
ulate the problem of automated discovery of diverse self-organized patterns in s
uch high-dimensional complex dynamical systems, as well as a framework for exper
imentation and evaluation. Using a continuous GOL as a testbed, we show that rec
ent intrinsically-motivated machine learning algorithms (POP-IMGEPs), initially
developed for learning of inverse models in robotics, can be transposed and used
 in this novel application area. These algorithms combine intrinsically-motivate
d goal exploration and unsupervised learning of goal space representations. Goal
 space representations describe the interesting features of patterns for which d
iverse variations should be discovered. In particular, we compare various approa
ches to define and learn goal space representations from the perspective of disc
overing diverse spatially localized patterns. Moreover, we introduce an extensio
n of a state-of-the-art POP-IMGEP algorithm which incrementally learns a goal re
presentation using a deep auto-encoder, and the use of CPPN primitives for gener
ating initialization parameters. We show that it is more efficient than several
baselines and equally efficient as a system pre-trained on a hand-made database
of patterns identified by human experts.
****
The Ingredients of Real World Robotic Reinforcement Learning
Henry Zhu,Justin Yu,Abhishek Gupta,Dhruv Shah,Kristian Hartikainen,Avi Singh,Vik
ash Kumar,Sergey Levine
The success of reinforcement learning in the real world has been limited to inst
rumented laboratory scenarios, often requiring arduous human supervision to enab

le continuous learning. In this work, we discuss the required elements of a robo tic system that can continually and autonomously improve with data collected in the real world, and propose a particular instantiation of such a system. Subsequ ently, we investigate a number of challenges of learning without instrumentation -- including the lack of episodic resets, state estimation, and hand-engineered rewards -- and propose simple, scalable solutions to these challenges. We demon strate the efficacy of our proposed system on dexterous robotic manipulation tas ks in simulation and the real world, and also provide an insightful analysis and ablation study of the challenges associated with this learning paradigm.
****
Causal Discovery with Reinforcement Learning
Shengyu Zhu,Ignavier Ng,Zhitang Chen
Discovering causal structure among a set of variables is a fundamental problem i n many empirical sciences. Traditional score-based casual discovery methods rely on various local heuristics to search for a Directed Acyclic Graph (DAG) accord ing to a predefined score function. While these methods, e.g., greedy equivalenc e search, may have attractive results with infinite samples and certain model as sumptions, they are less satisfactory in practice due to finite data and possibl e violation of assumptions. Motivated by recent advances in neural combinatorial optimization, we propose to use Reinforcement Learning (RL) to search for the D AG with the best scoring. Our encoder-decoder model takes observable data as inp ut and generates graph adjacency matrices that are used to compute rewards. The reward incorporates both the predefined score function and two penalty terms for enforcing acyclicity. In contrast with typical RL applications where the goal i s to learn a policy, we use RL as a search strategy and our final output would b e the graph, among all graphs generated during training, that achieves the best reward. We conduct experiments on both synthetic and real datasets, and show tha t the proposed approach not only has an improved search ability but also allows for a flexible score function under the acyclicity constraint.
****
Modelling the influence of data structure on learning in neural networks
S. Goldt,M. Mézard,F. Krzakala,L. Zdeborová
The lack of crisp mathematical models that capture the structure of real-world data sets is a major obstacle to the detailed theoretical understanding of deep neural networks. Here, we first demonstrate the effect of structured data sets by experimentally comparing the dynamics and the performance of two-layer networks trained on two different data sets: (i) an unstructured synthetic data set containing random i.i.d. inputs, and (ii) a simple canonical data set such as MNIST images. Our analysis reveals two phenomena related to the dynamics of the networks and their ability to generalise that only appear when training on structured data sets. Second, we introduce a generative model for data sets, where high-dimensional inputs lie on a lower-dimensional manifold and have labels that depend only on their position within this manifold. We call it the *hidden manifold model* and we experimentally demonstrate that training networks on data sets drawn from this model reproduces both the phenomena seen during training on MNIST.
****
Scaling Autoregressive Video Models
Dirk Weissenborn,Oscar Täckström,Jakob Uszkoreit
Due to the statistical complexity of video, the high degree of inherent stochast icity, and the sheer amount of data, generating natural video remains a challeng ing task. State-of-the-art video generation models attempt to address these issu es by combining sometimes complex, often video-specific neural network architect ures, latent variable models, adversarial training and a range of other methods. Despite their often high complexity, these approaches still fall short of gener ating high quality video continuations outside of narrow domains and often strug gle with fidelity. In contrast, we show that conceptually simple, autoregressive video generation models based on a three-dimensional self-attention mechanism a chieve highly competitive results across multiple metrics on popular benchmark d atasets for which they produce continuations of high fidelity and realism. Furth

ermore, we find that our models are capable of producing diverse and surprisingl
y realistic continuations on a subset of videos from Kinetics, a large scale act
ion recognition dataset comprised of YouTube videos exhibiting phenomena such as
 camera movement, complex object interactions and diverse human movement. To our
 knowledge, this is the first promising application of video-generation models t
o videos of this complexity.
****

TOWARDS FEATURE SPACE ADVERSARIAL ATTACK
Qiuling Xu,Guanhong Tao,Siyuan Cheng,Lin Tan,Xiangyu Zhang
We propose a new type of adversarial attack to Deep Neural Networks (DNNs) for i
mage classification. Different from most existing attacks that directly perturb
input pixels. Our attack focuses on perturbing abstract features, more specifica
lly, features that denote styles, including interpretable styles such as vivid c
olors and sharp outlines, and uninterpretable ones. It induces model misclassfic
ation by injecting style changes insensitive for humans, through an optimization
 procedure. We show that state-of-the-art pixel space adversarial attack detecti
on and defense techniques are ineffective in guarding against feature space atta
cks.
****

Convergence Analysis of a Momentum Algorithm with Adaptive Step Size for Nonconv
ex Optimization
Anas Barakat,Pascal Bianchi
Although Adam is a very popular algorithm for optimizing the weights of neural n
etworks, it has been recently shown that it can diverge even in simple convex op
timization examples. Therefore, several variants of Adam have been proposed to c
ircumvent this convergence issue. In this work, we study the algorithm for smoot
h nonconvex optimization under a boundedness assumption on the adaptive learning
 rate. The bound on the adaptive step size depends on the Lipschitz constant of
the gradient of the objective function and provides safe theoretical adaptive st
ep sizes. Under this boundedness assumption, we show a novel first order converg
ence rate result in both deterministic and stochastic contexts. Furthermore, we
establish convergence rates of the function value sequence using the Kurdyka-Loj
asiewicz property.
****

Compressive Transformers for Long-Range Sequence Modelling
Jack W. Rae,Anna Potapenko,Siddhant M. Jayakumar,Chloe Hillier,Timothy P. Lillic
rap
We present the Compressive Transformer, an attentive sequence model which compre
sses past memories for long-range sequence learning. We find the Compressive Tra
nsformer obtains state-of-the-art language modelling results in the WikiText-103
 and Enwik8 benchmarks, achieving 17.1 ppl and 0.97bpc respectively. We also fin
d it can model high-frequency speech effectively and can be used as a memory mec
hanism for RL, demonstrated on an object matching task. To promote the domain of
 long-range sequence learning, we propose a new open-vocabulary language modelli
ng benchmark derived from books, PG-19.
****

Global Momentum Compression for Sparse Communication in Distributed SGD
Shen-Yi Zhao,Yin-Peng Xie,Hao Gao,Wu-Jun Li
With the rapid growth of data, distributed stochastic gradient descent~(DSGD) ha
s been widely used for solving large-scale machine learning problems. Due to the
 latency and limited bandwidth of network, communication has become the bottlene
ck of DSGD when we need to train large scale models, like deep neural networks.
Communication compression with sparsified gradient, abbreviated as \emph{sparse
communication}, has been widely used for reducing communication cost in DSGD. Re
cently, there has appeared one method, called deep gradient compression~(DGC), t
o combine memory gradient and momentum SGD for sparse communication. DGC has ach
ieved promising performance in practice. However, the theory about the convergen
ce of DGC is lack. In this paper, we propose a novel method, called \emph{\under
line{g}}lobal \emph{\underline{m}}omentum \emph{\underline{c}}ompression~(GMC),
for sparse communication in DSGD. GMC also combines memory gradient and momentum

SGD. But different from DGC which adopts local momentum, GMC adopts global momentum. We theoretically prove the convergence rate of GMC for both convex and non-convex problems. To the best of our knowledge, this is the first work that proves the convergence of distributed momentum SGD~(DMSGD) with sparse communication and memory gradient. Empirical results show that, compared with the DMSGD counterpart without sparse communication, GMC can reduce the communication cost by approximately 100 fold without loss of generalization accuracy. GMC can also achieve comparable~(sometimes better) performance compared with DGC, with an extra theoretical guarantee.

****

Differentiation of Blackbox Combinatorial Solvers

Marin Vlastelica Pogan■i■,Anselm Paulus,Vit Musil,Georg Martius,Michal Rolinek

Achieving fusion of deep learning with combinatorial algorithms promises transformative changes to artificial intelligence. One possible approach is to introduce combinatorial building blocks into neural networks. Such end-to-end architectures have the potential to tackle combinatorial problems on raw input data such as ensuring global consistency in multi-object tracking or route planning on maps in robotics. In this work, we present a method that implements an efficient backward pass through blackbox implementations of combinatorial solvers with linear objective functions. We provide both theoretical and experimental backing. In particular, we incorporate the Gurobi MIP solver, Blossom V algorithm, and Dijkstra's algorithm into architectures that extract suitable features from raw inputs for the traveling salesman problem, the min-cost perfect matching problem and the shortest path problem.

****

Reinforced Genetic Algorithm Learning for Optimizing Computation Graphs

Aditya Paliwal,Felix Gimeno,Vinod Nair,Yujia Li,Miles Lubin,Pushmeet Kohli,Oriol Vinyals

We present a deep reinforcement learning approach to minimizing the execution cost of neural network computation graphs in an optimizing compiler. Unlike earlier learning-based works that require training the optimizer on the same graph to be optimized, we propose a learning approach that trains an optimizer offline and then generalizes to previously unseen graphs without further training. This allows our approach to produce high-quality execution decisions on real-world TensorFlow graphs in seconds instead of hours. We consider two optimization tasks for computation graphs: minimizing running time and peak memory usage. In comparison to an extensive set of baselines, our approach achieves significant improvements over classical and other learning-based methods on these two tasks.

****

Lagrangian Fluid Simulation with Continuous Convolutions

Benjamin Ummenhofer,Lukas Prantl,Nils Thuerey,Vladlen Koltun

We present an approach to Lagrangian fluid simulation with a new type of convolutional network. Our networks process sets of moving particles, which describe fluids in space and time. Unlike previous approaches, we do not build an explicit graph structure to connect the particles but use spatial convolutions as the main differentiable operation that relates particles to their neighbors. To this end we present a simple, novel, and effective extension of N-D convolutions to the continuous domain. We show that our network architecture can simulate different materials, generalizes to arbitrary collision geometries, and can be used for inverse problems. In addition, we demonstrate that our continuous convolutions outperform prior formulations in terms of accuracy and speed.

****

Harnessing the Power of Infinitely Wide Deep Nets on Small-data Tasks

Sanjeev Arora,Simon S. Du,Zhiyuan Li,Ruslan Salakhutdinov,Ruosong Wang,Dingli Yu

Recent research shows that the following two models are equivalent: (a) infinitely wide neural networks (NNs) trained under l2 loss by gradient descent with infinitesimally small learning rate (b) kernel regression with respect to so-called Neural Tangent Kernels (NTKs) (Jacot et al., 2018). An efficient algorithm to compute the NTK, as well as its convolutional counterparts, appears in Arora et a

l. (2019a), which allowed studying performance of infinitely wide nets on datase
ts like CIFAR-10. However, super-quadratic running time of kernel methods makes
them best suited for small-data tasks. We report results suggesting neural tange
nt kernels perform strongly on low-data tasks.
1. On a standard testbed of classification/regression tasks from the UCI databas
e, NTK SVM beats the previous gold standard, Random Forests (RF), and also the c
orresponding finite nets.
2. On CIFAR-10 with 10 – 640 training samples, Convolutional NTK consistently be
ats ResNet-34 by 1% - 3%.
3. On VOC07 testbed for few-shot image classification tasks on ImageNet with tra
nsfer learning (Goyal et al., 2019), replacing the linear SVM currently used wit
h a Convolutional NTK SVM consistently improves performance.
4. Comparing the performance of NTK with the finite-width net it was derived fro
m, NTK behavior starts at lower net widths than suggested by theoretical analysi
s(Arora et al., 2019a). NTK’s efficacy may trace to lower variance of output.
****

Semi-supervised semantic segmentation needs strong, high-dimensional perturbatio
ns
Geoff French,Timo Aila,Samuli Laine,Michal Mackiewicz,Graham Finlayson
Consistency regularization describes a class of approaches that have yielded gro
und breaking results in semi-supervised classification problems. Prior work has
established the cluster assumption\,---\,under which the data distribution consi
sts of uniform class clusters of samples separated by low density regions\,---\,
as key to its success. We analyze the problem of semantic segmentation and find
that the data distribution does not exhibit low density regions separating class
es and offer this as an explanation for why semi-supervised segmentation is a ch
allenging problem.
We then identify the conditions that allow consistency regularization to work ev
en without such low-density regions.
This allows us to generalize the recently proposed CutMix augmentation technique
 to a powerful masked variant, CowMix,
leading to a successful application of consistency regularization in the semi-su
pervised semantic segmentation setting and
reaching state-of-the-art results in several standard datasets.
****

Learning to Guide Random Search
Ozan Sener,Vladlen Koltun
We are interested in derivative-free optimization of high-dimensional functions.
 The sample complexity of existing methods is high and depends on problem dimens
ionality, unlike the dimensionality-independent rates of first-order methods. Th
e recent success of deep learning suggests that many datasets lie on low-dimensi
onal manifolds that can be represented by deep nonlinear models. We therefore co
nsider derivative-free optimization of a high-dimensional function that lies on
a latent low-dimensional manifold. We develop an online learning approach that l
earns this manifold while performing the optimization. In other words, we jointl
y learn the manifold and optimize the function. Our analysis suggests that the p
resented method significantly reduces sample complexity. We empirically evaluate
 the method on continuous optimization benchmarks and high-dimensional continuou
s control problems. Our method achieves significantly lower sample complexity th
an Augmented Random Search, Bayesian optimization, covariance matrix adaptation
(CMA-ES), and other derivative-free optimization algorithms.
****

Attentive Sequential Neural Processes
Jaesik Yoon,Gautam Singh,Sungjin Ahn
Sequential Neural Processes (SNP) is a new class of models that can meta-learn a
 temporal stochastic process of stochastic processes by modeling temporal transi
tion between Neural Processes. As Neural Processes (NP) suffers from underfittin
g, SNP is also prone to the same problem, even more severely due to its temporal
 context compression. Applying attention which resolves the problem of NP, howev
er, is a challenge in SNP, because it cannot store the past contexts over which

it is supposed to apply attention. In this paper, we propose the Attentive Sequential Neural Processes (ASNP) that resolve the underfitting in SNP by introducing a novel imaginary context as a latent variable and by applying attention over the imaginary context. We evaluate our model on 1D Gaussian Process regression and 2D moving MNIST/CelebA regression. We apply ASNP to implement Attentive Temporal GQN and evaluate on the moving-CelebA task.
****
The intriguing role of module criticality in the generalization of deep networks
Niladri Chatterji,Behnam Neyshabur,Hanie Sedghi
We study the phenomenon that some modules of deep neural networks (DNNs) are more critical than others. Meaning that rewinding their parameter values back to initialization, while keeping other modules fixed at the trained parameters, results in a large drop in the network's performance. Our analysis reveals interesting properties of the loss landscape which leads us to propose a complexity measure, called module criticality, based on the shape of the valleys that connect the initial and final values of the module parameters. We formulate how generalization relates to the module criticality, and show that this measure is able to explain the superior generalization performance of some architectures over others, whereas, earlier measures fail to do so.
****
Yet another but more efficient black-box adversarial attack: tiling and evolution strategies
Laurent Meunier,Jamal Atif,Olivier Teytaud
We introduce a new black-box attack achieving state of the art performances. Our approach is based on a new objective function, borrowing ideas from ℓ_∞-white box attacks, and particularly designed to fit derivative-free optimization requirements. It only requires to have access to the logits of the classifier without any other information which is a more realistic scenario. Not only we introduce a new objective function, we extend previous works on black box adversarial attacks to a larger spectrum of evolution strategies and other derivative-free optimization methods. We also highlight a new intriguing property that deep neural networks are not robust to single shot tiled attacks. Our models achieve, with a budget limited to $10,000$ queries, results up to 99.2% of success rate against InceptionV3 classifier with 630 queries to the network on average in the untargeted attacks setting, which is an improvement by 90 queries of the current state of the art. In the targeted setting, we are able to reach, with a limited budget of $100,000$, 100% of success rate with a budget of $6,662$ queries on average, i.e. we need 800 queries less than the current state of the art.
****
TreeCaps: Tree-Structured Capsule Networks for Program Source Code Processing
Vinoj Jayasundara,Nghi Duy Quoc Bui,Lingxiao Jiang,David Lo
Program comprehension is a fundamental task in software development and maintenance processes. Software developers often need to understand a large amount of existing code before they can develop new features or fix bugs in existing programs. Being able to process programming language code automatically and provide summaries of code functionality accurately can significantly help developers to reduce time spent in code navigation and understanding, and thus increase productivity. Different from natural language articles, source code in programming languages often follows rigid syntactical structures and there can exist dependencies among code elements that are located far away from each other through complex control flows and data flows. Existing studies on tree-based convolutional neural networks (TBCNN) and gated graph neural networks (GGNN) are not able to capture essential semantic dependencies among code elements accurately. In this paper, we propose novel tree-based capsule networks (TreeCaps) and relevant techniques for processing program code in an automated way that encodes code syntactical structures and captures code dependencies more accurately. Based on evaluation on programs written in different programming languages, we show that our TreeCaps-based approach can outperform other approaches in classifying the functionalities of many programs.

****
Learning with Social Influence through Interior Policy Differentiation
Hao Sun,Bo Dai,Jiankai Sun,Zhenghao Peng,Guodong Xu,Dahua Lin,Bolei Zhou
Animals develop novel skills not only through the interaction with the environment but also from the influence of the others. In this work we model the social influence into the scheme of reinforcement learning, enabling the agents to learn both from the environment and from their peers. Specifically, we first define a metric to measure the distance between policies then quantitatively derive the definition of uniqueness. Unlike previous precarious joint optimization approaches, the social uniqueness motivation in our work is imposed as a constraint to encourage the agent to learn a policy different from the existing agents while still solve the primal task. The resulting algorithm, namely Interior Policy Differentiation (IPD), brings about performance improvement as well as a collection of policies that solve a given task with distinct behaviors
****
SPROUT: Self-Progressing Robust Training
Minhao Cheng,Pin-Yu Chen,Sijia Liu,Shiyu Chang,Cho-Jui Hsieh,Payel Das
Enhancing model robustness under new and even adversarial environments is a crucial milestone toward building trustworthy and reliable machine learning systems. Current robust training methods such as adversarial training explicitly specify an ``attack'' (e.g., ℓ_{∞}-norm bounded perturbation) to generate adversarial examples during model training in order to improve adversarial robustness. In this paper, we take a different perspective and propose a new framework SPROUT, self-progressing robust training. During model training, SPROUT progressively adjusts training label distribution via our proposed parametrized label smoothing technique, making training free of attack generation and more scalable. We also motivate SPROUT using a general formulation based on vicinity risk minimization, which includes many robust training methods as special cases. Compared with state-of-the-art adversarial training methods (PGD-ℓ_∞ and TRADES) under ℓ_{∞}-norm bounded attacks and various invariance tests, SPROUT consistently attains superior performance and is more scalable to large neural networks. Our results shed new light on scalable, effective and attack-independent robust training methods.
****
Alleviating Privacy Attacks via Causal Learning
Shruti Tople,Amit Sharma,Aditya Nori
Machine learning models, especially deep neural networks have been shown to reveal membership information of inputs in the training data. Such membership inference attacks are a serious privacy concern, for example, patients providing medical records to build a model that detects HIV would not want their identity to be leaked. Further, we show that the attack accuracy amplifies when the model is used to predict samples that come from a different distribution than the training set, which is often the case in real world applications. Therefore, we propose the use of causal learning approaches where a model learns the causal relationship between the input features and the outcome. Causal models are known to be invariant to the training distribution and hence generalize well to shifts between samples from the same distribution and across different distributions. First, we prove that models learned using causal structure provide stronger differential privacy guarantees than associational models under reasonable assumptions. Next, we show that causal models trained on sufficiently large samples are robust to membership inference attacks across different distributions of datasets and those trained on smaller sample sizes always have lower attack accuracy than corresponding associational models. Finally, we confirm our theoretical claims with experimental evaluation on 4 datasets with moderately complex Bayesian networks. We observe that neural network-based associational models exhibit upto 80% attack accuracy under different test distributions and sample sizes whereas causal models exhibit attack accuracy close to a random guess. Our results confirm the value of the generalizability of causal models in reducing susceptibility to privacy attacks.
****

Hybrid Weight Representation: A Quantization Method Represented with Ternary and Sparse-Large Weights
Jinbae Park,Sung-Ho Bae
Previous ternarizations such as the trained ternary quantization (TTQ), which quantized weights to three values (e.g., {−Wn, 0,+Wp}), achieved the small model size and efficient inference process. However, the extreme limit on the number of quantization steps causes some degradation in accuracy. To solve this problem, we propose a hybrid weight representation (HWR) method which produces a network consisting of two types of weights, i.e., ternary weights (TW) and sparse-large weights (SLW). The TW is similar to the TTQ's and requires three states to be stored in memory with 2 bits. We utilize the one remaining state to indicate the SLW which is referred to as very rare and greater than TW. In HWR, we represent TW with values while SLW with indices of values. By encoding SLW, the networks can preserve their model size with improving their accuracy. To fully utilize HWR, we also introduce a centralized quantization (CQ) process with a weighted ridge (WR) regularizer. They aim to reduce the entropy of weight distributions by centralizing weights toward ternary values. Our comprehensive experiments show that HWR outperforms the state-of-the-art compressed models in terms of the trade-off between model size and accuracy. Our proposed representation increased the AlexNet performance on CIFAR-100 by 4.15% with only1.13% increase in model size.
****
Self-labelling via simultaneous clustering and representation learning
Asano YM.,Rupprecht C.,Vedaldi A.
Combining clustering and representation learning is one of the most promising approaches for unsupervised learning of deep neural networks. However, doing so naively leads to ill posed learning problems with degenerate solutions.
In this paper, we propose a novel and principled learning formulation that addresses these issues.
The method is obtained by maximizing the information between labels and input data indices.
We show that this criterion extends standard cross-entropy minimization to an optimal transport problem, which we solve efficiently for millions of input images and thousands of labels using a fast variant of the Sinkhorn-Knopp algorithm.
The resulting method is able to self-label visual data so as to train highly competitive image representations without manual labels. Our method achieves state of the art representation learning performance for AlexNet and ResNet-50 on SVHN, CIFAR-10, CIFAR-100 and ImageNet and yields the first self-supervised AlexNet that outperforms the supervised Pascal VOC detection baseline.
****
Continual Learning with Gated Incremental Memories for Sequential Data Processing
Andrea Cossu,Antonio Carta,Davide Bacciu
The ability to learn over changing task distributions without forgetting previous knowledge, also known as continual learning, is a key enabler for scalable and trustworthy deployments of adaptive solutions. While the importance of continual learning is largely acknowledged in machine vision and reinforcement learning problems, this is mostly under-documented for sequence processing tasks. This work focuses on characterizing and quantitatively assessing the impact of catastrophic forgetting and task interference when dealing with sequential data in recurrent neural networks. We also introduce a general architecture, named Gated Incremental Memory, for augmenting recurrent models with continual learning skills, whose effectiveness is demonstrated through the benchmarks introduced in this paper.
****
Policy Optimization by Local Improvement through Search
Jialin Song,Joe Wenjie Jiang,Amir Yazdanbakhsh,Ebrahim Songhori,Anna Goldie,Navdeep Jaitly,Azalia Mirhoseini
Imitation learning has emerged as a powerful strategy for learning initial policies that can be refined with reinforcement learning techniques. Most strategies

in imitation learning, however, rely on per-step supervision either from expert demonstrations, referred to as behavioral cloning or from interactive expert policy queries such as DAgger. These strategies differ on the state distribution at which the expert actions are collected -- the former using the state distribution of the expert, the latter using the state distribution of the policy being trained. However, the learning signal in both cases arises from the expert actions. On the other end of the spectrum, approaches rooted in Policy Iteration, such as Dual Policy Iteration do not choose next step actions based on an expert, but instead use planning or search over the policy to choose an action distribution to train towards. However, this can be computationally expensive, and can also end up training the policy on a state distribution that is far from the current policy's induced distribution. In this paper, we propose an algorithm that finds a middle ground by using Monte Carlo Tree Search (MCTS) to perform local trajectory improvement over rollouts from the policy. We provide theoretical justification for both the proposed local trajectory search algorithm and for our use of MCTS as a local policy improvement operator. We also show empirically that our method (Policy Optimization by Local Improvement through Search or POLISH) is much faster than methods that plan globally, speeding up training by a factor of up to 14 in wall clock time. Furthermore, the resulting policy outperforms strong baselines in both reinforcement learning and imitation learning.
****

Improving Model Compatibility of Generative Adversarial Networks by Boundary Calibration

Si-An Chen,Chun-Liang Li,Hsuan-Tien Lin

Generative Adversarial Networks (GANs) is a powerful family of models that learn an underlying distribution to generate synthetic data. Many existing studies of GANs focus on improving the realness of the generated image data for visual applications, and few of them concern about improving the quality of the generated data for training other classifiers---a task known as the model compatibility problem. As a consequence, existing GANs often prefer generating `easier' synthetic data that are far from the boundaries of the classifiers, and refrain from generating near-boundary data, which are known to play an important roles in training the classifiers. To improve GAN in terms of model compatibility, we propose Boundary-Calibration GANs (BCGANs), which leverage the boundary information from a set of pre-trained classifiers using the original data. In particular, we introduce an auxiliary Boundary-Calibration loss (BC-loss) into the generator of GAN to match the statistics between the posterior distributions of original data and generated data with respect to the boundaries of the pre-trained classifiers. The BC-loss is provably unbiased and can be easily coupled with different GAN variants to improve their model compatibility. Experimental results demonstrate that BCGANs not only generate realistic images like original GANs but also achieves superior model compatibility than the original GANs.
****

Robust anomaly detection and backdoor attack detection via differential privacy

Min Du,Ruoxi Jia,Dawn Song

Outlier detection and novelty detection are two important topics for anomaly detection. Suppose the majority of a dataset are drawn from a certain distribution, outlier detection and novelty detection both aim to detect data samples that do not fit the distribution. Outliers refer to data samples within this dataset, while novelties refer to new samples. In the meantime, backdoor poisoning attacks for machine learning models are achieved through injecting poisoning samples into the training dataset, which could be regarded as “outliers” that are intentionally added by attackers. Differential privacy has been proposed to avoid leaking any individual’s information, when aggregated analysis is performed on a given dataset. It is typically achieved by adding random noise, either directly to the input dataset, or to intermediate results of the aggregation mechanism. In this paper, we demonstrate that applying differential privacy could improve the utility of outlier detection and novelty detection, with an extension to detect poisoning samples in backdoor attacks. We first present a theoretical analysis on how differential privacy helps with the detection, and then conduct extensive exp

eriments to validate the effectiveness of differential privacy in improving outl
ier detection, novelty detection, and backdoor attack detection.
****
CAT: Compression-Aware Training for bandwidth reduction
Chaim Baskin,Brian Chmiel,Evgenii Zheltonozhskii,Ron Banner,Alex M. Bronstein,Av
i Mendelson
Convolutional neural networks (CNNs) have become the dominant neural network arc
hitecture for solving visual processing tasks. One of the major obstacles hinder
ing the ubiquitous use of CNNs for inference is their relatively high memory ban
dwidth requirements, which can be a main energy consumer and throughput bottlene
ck in hardware accelerators. Accordingly, an efficient feature map compression m
ethod can result in substantial performance gains. Inspired by quantization-awar
e training approaches, we propose a compression-aware training (CAT) method that
 involves training the model in a way that allows better compression of feature
maps during inference. Our method trains the model to achieve low-entropy featur
e maps, which enables efficient compression at inference time using classical tr
ansform coding methods. CAT significantly improves the state-of-the-art results
reported for quantization. For example, on ResNet-34 we achieve 73.1% accuracy (
0.2% degradation from the baseline) with an average representation of only 1.79
 bits per value. Reference implementation accompanies the paper.
****
Scheduling the Learning Rate Via Hypergradients: New Insights and a New Algorith
m
Michele Donini,Luca Franceschi,Orchid Majumder,Massimiliano Pontil,Paolo Frascon
i
We study the problem of fitting task-specific learning rate schedules from the p
erspective of hyperparameter optimization. This allows us to explicitly search
for schedules that achieve good generalization. We describe the structure of the
 gradient of a validation error w.r.t. the learning rates, the hypergradient, an
d based on this we introduce a novel online algorithm. Our method adaptively int
erpolates between two recently proposed techniques (Franceschi et al., 2017; Bay
din et al.,2018), featuring increased stability and faster convergence. We show
empirically that the proposed technique compares favorably with baselines and re
lated methodsin terms of final test accuracy.
****
Learning Entailment-Based Sentence Embeddings from Natural Language Inference
Rabeeh Karimi Mahabadi*,Florian Mai*,James Henderson
Large datasets on natural language inference are a potentially valuable resource
 for inducing semantic representations of natural language sentences. But in ma
ny such models the embeddings computed by the sentence encoder goes through an M
LP-based interaction layer before predicting its label, and thus some of the inf
ormation about textual entailment is encoded in the interpretation of sentence e
mbeddings given by this parameterised MLP.
In this work we propose a simple interaction layer based on predefined entailmen
t and contradiction scores applied directly to the sentence embeddings. This pa
rameter-free interaction model achieves results on natural language inference co
mpetitive with MLP-based models, demonstrating that the trained sentence embeddi
ngs directly represent the information needed for textual entailment, and the in
ductive bias of this model leads to better generalisation to other related datas
ets.
****
Invariance vs Robustness of Neural Networks
Sandesh Kamath,Amit Deshpande,K V Subrahmanyam
Neural networks achieve human-level accuracy on many standard datasets used in i
mage classification. The next step is to achieve better generalization to natura
l (or non-adversarial) perturbations as well as known pixel-wise adversarial per
turbations of inputs. Previous work has studied generalization to natural geomet
ric transformations (e.g., rotations) as invariance, and generalization to adver
sarial perturbations as robustness. In this paper, we examine the interplay betw
een invariance and robustness. We empirically study the following two cases:(a)

change in adversarial robustness as we improve only the invariance using equivariant models and training augmentation, (b) change in invariance as we improve only the adversarial robustness using adversarial training. We observe that the rotation invariance of equivariant models (StdCNNs and GCNNs) improves by training augmentation with progressively larger rotations but while doing so, their adversarial robustness does not improve, or worse, it can even drop significantly on datasets such as MNIST. As a plausible explanation for this phenomenon we observe that the average perturbation distance of the test points to the decision boundary decreases as the model learns larger and larger rotations. On the other hand, we take adversarially trained LeNet and ResNet models which have good \ell_\infty adversarial robustness on MNIST and CIFAR-10, and observe that adversarially training them with progressively larger norms keeps their rotation invariance essentially unchanged. In fact, the difference between test accuracy on unrotated test data and on randomly rotated test data upto \theta , for all \theta in [0, 180], remains essentially unchanged after adversarial training . As a plausible explanation for the observed phenomenon we show empirically that the principal components of adversarial perturbations and perturbations given by small rotations are nearly orthogonal

****

Asymptotic learning curves of kernel methods: empirical data v.s. Teacher-Student paradigm

Stefano Spigler,Mario Geiger,Matthieu Wyart

How many training data are needed to learn a supervised task? It is often observed that the generalization error decreases as $n^{-\beta}$ where n is the number of training examples and β an exponent that depends on both data and algorithm. In this work we measure β when applying kernel methods to real datasets. For MNIST we find $\beta\approx 0.4$ and for CIFAR10 $\beta\approx 0.1$. Remarkably, β is the same for regression and classification tasks, and for Gaussian or Laplace kernels. To rationalize the existence of non-trivial exponents that can be independent of the specific kernel used, we introduce the Teacher-Student framework for kernels. In this scheme, a Teacher generates data according to a Gaussian random field, and a Student learns them via kernel regression. With a simplifying assumption --- namely that the data are sampled from a regular lattice --- we derive analytically β for translation invariant kernels, using previous results from the kriging literature. Provided that the Student is not too sensitive to high frequencies, β depends only on the training data and their dimension. We confirm numerically that these predictions hold when the training points are sampled at random on a hypersphere. Overall, our results quantify how smooth Gaussian data should be to avoid the curse of dimensionality, and indicate that for kernel learning the relevant dimension of the data should be defined in terms of how the distance between nearest data points depends on n. With this definition one obtains reasonable effective smoothness estimates for MNIST and CIFAR10.

****

LARGE SCALE REPRESENTATION LEARNING FROM TRIPLET COMPARISONS

Siavash Haghiri,Leena Chennuru Vankadara,Ulrike von Luxburg

In this paper, we discuss the fundamental problem of representation learning from a new perspective. It has been observed in many supervised/unsupervised DNNs that the final layer of the network often provides an informative representation for many tasks, even though the network has been trained to perform a particular task. The common ingredient in all previous studies is a low-level feature representation for items, for example, RGB values of images in the image context. In the present work, we assume that no meaningful representation of the items is given. Instead, we are provided with the answers to some triplet comparisons of the following form: Is item A more similar to item B or item C? We provide a fast algorithm based on DNNs that constructs a Euclidean representation for the items, using solely the answers to the above-mentioned triplet comparisons. This problem has been studied in a sub-community of machine learning by the name "Ordinal Embedding". Previous approaches to the problem are painfully slow and cannot scale to larger datasets. We demonstrate that our proposed approach is significan

tly faster than available methods, and can scale to real-world large datasets.

Thereby, we also draw attention to the less explored idea of using neural networ
ks to directly, approximately solve non-convex, NP-hard optimization problems th
at arise naturally in unsupervised learning problems.
****
Learning to Reach Goals Without Reinforcement Learning
Dibya Ghosh,Abhishek Gupta,Justin Fu,Ashwin Reddy,Coline Devin,Benjamin Eysenbac
h,Sergey Levine
Imitation learning algorithms provide a simple and straightforward approach for
training control policies via standard supervised learning methods. By maximizin
g the likelihood of good actions provided by an expert demonstrator, supervised
imitation learning can produce effective policies without the algorithmic comple
xities and optimization challenges of reinforcement learning, at the cost of req
uiring an expert demonstrator -- typically a person -- to provide the demonstrat
ions. In this paper, we ask: can we use imitation learning to train effective po
licies without any expert demonstrations? The key observation that makes this po
ssible is that, in the multi-task setting, trajectories that are generated by a
suboptimal policy can still serve as optimal examples for other tasks. In partic
ular, in the setting where the tasks correspond to different goals, every trajec
tory is a successful demonstration for the state that it actually reaches. Infor
med by this observation, we propose a very simple algorithm for learning behavio
rs without any demonstrations, user-provided reward functions, or complex reinfo
rcement learning methods. Our method simply maximizes the likelihood of actions
the agent actually took in its own previous rollouts, conditioned on the goal be
ing the state that it actually reached. Although related variants of this approa
ch have been proposed previously in imitation learning settings with example dem
onstrations, we present the first instance of this approach as a method for lear
ning goal-reaching policies entirely from scratch. We present a theoretical resu
lt linking self-supervised imitation learning and reinforcement learning, and em
pirical results showing that it performs competitively with more complex reinfor
cement learning methods on a range of challenging goal reaching problems.
****
Subjective Reinforcement Learning for Open Complex Environments
Zhile Yang*,Haichuan Gao*,Xin Su,Shangqi Guo,Feng Chen
Solving tasks in open environments has been one of the long-time pursuits of rei
nforcement learning researches. We propose that data confusion is the core under
lying problem. Although there exist methods that implicitly alleviate it from di
fferent perspectives, we argue that their solutions are based on task-specific p
rior knowledge that is constrained to certain kinds of tasks and lacks theoretic
al guarantees. In this paper, Subjective Reinforcement Learning Framework is pro
posed to state the problem from a broader and systematic view, and subjective po
licy is proposed to represent existing related algorithms in general. Theoretica
l analysis is given about the conditions for the superiority of a subjective pol
icy, and the relationship between model complexity and the overall performance.
Results are further applied as guidance for algorithm designing without task-spe
cific prior knowledge about tasks.

****
Deep probabilistic subsampling for task-adaptive compressed sensing
Iris A.M. Huijben,Bastiaan S. Veeling,Ruud J.G. van Sloun
The field of deep learning is commonly concerned with optimizing predictive mode
ls using large pre-acquired datasets of densely sampled datapoints or signals. I
n this work, we demonstrate that the deep learning paradigm can be extended to i
ncorporate a subsampling scheme that is jointly optimized under a desired minimu
m sample rate. We present Deep Probabilistic Subsampling (DPS), a widely applica
ble framework for task-adaptive compressed sensing that enables end-to end optim
ization of an optimal subset of signal samples with a subsequent model that perf
orms a required task. We demonstrate strong performance on reconstruction and cl
assification tasks of a toy dataset, MNIST, and CIFAR10 under stringent subsampl

ing rates in both the pixel and the spatial frequency domain. Due to the task-ag
nostic nature of the framework, DPS is directly applicable to all real-world dom
ains that benefit from sample rate reduction.
****

Semi-supervised 3D Face Reconstruction with Nonlinear Disentangled Representatio
ns

Zhongpai Gao,Juyong Zhang,Yudong Guo,Chao Ma,Guangtao Zhai,Xiaokang Yang

Recovering 3D geometry shape, albedo and lighting from a single image has wide a
pplications in many areas, which is also a typical ill-posed problem. In order t
o eliminate the ambiguity, face prior knowledge like linear 3D morphable models
(3DMM) learned from limited scan data are often adopted to the reconstruction pr
ocess. However, methods based on linear parametric models cannot generalize well
 for facial images in the wild with various ages, ethnicity, expressions, poses,
 and lightings. Recent methods aim to learn a nonlinear parametric model using c
onvolutional neural networks (CNN) to regress the face shape and texture directl
y. However, the models were only trained on a dataset that is generated from a l
inear 3DMM. Moreover, the identity and expression representations are entangled
in these models, which hurdles many facial editing applications. In this paper,
we train our model with adversarial loss in a semi-supervised manner on hybrid b
atches of unlabeled and labeled face images to exploit the value of large amount
s of unlabeled face images from unconstrained photo collections. A novel center
loss is introduced to make sure that different facial images from the same perso
n have the same identity shape and albedo. Besides, our proposed model disentang
les identity, expression, pose, and lighting representations, which improves the
 overall reconstruction performance and facilitates facial editing applications,
 e.g., expression transfer. Comprehensive experiments demonstrate that our model
 produces high-quality reconstruction compared to state-of-the-art methods and i
s robust to various expression, pose, and lighting conditions.

****

Representing Model Uncertainty of Neural Networks in Sparse Information Form

Jongseok Lee,Rudolph Triebel

This paper addresses the problem of representing a system's belief using multi-v
ariate normal distributions (MND) where the underlying model is based on a deep
neural network (DNN). The major challenge with DNNs is the computational complex
ity that is needed to obtain model uncertainty using MNDs. To achieve a scalable
 method, we propose a novel approach that expresses the parameter posterior in s
parse information form. Our inference algorithm is based on a novel Laplace Appr
oximation scheme, which involves a diagonal correction of the Kronecker-factored
 eigenbasis. As this makes the inversion of the information matrix intractable -
 an operation that is required for full Bayesian analysis, we devise a low-rank
  approximation of this eigenbasis and a memory-efficient sampling scheme. We pr
ovide both a theoretical analysis and an empirical evaluation on various benchma
rk data sets, showing the superiority of our approach over existing methods.
****

GroSS Decomposition: Group-Size Series Decomposition for Whole Search-Space Trai
ning

Henry Howard-Jenkins,Yiwen Li,Victor Adrian Prisacariu

We present Group-size Series (GroSS) decomposition, a mathematical formulation o
f tensor factorisation into a series of approximations of increasing rank terms.
 GroSS allows for dynamic and differentiable selection of factorisation rank, wh
ich is analogous to a grouped convolution. Therefore, to the best of our knowled
ge, GroSS is the first method to simultaneously train differing numbers of group
s within a single layer, as well as all possible combinations between layers. In
 doing so, GroSS trains an entire grouped convolution architecture search-space
concurrently. We demonstrate this with a proof-of-concept exhaustive architecure
 search with a performance objective. GroSS represents a significant step toward
s liberating network architecture search from the burden of training and finetun
ing.
****

Neural Tangents: Fast and Easy Infinite Neural Networks in Python
Roman Novak,Lechao Xiao,Jiri Hron,Jaehoon Lee,Alexander A. Alemi,Jascha Sohl-Dickstein,Samuel S. Schoenholz
Neural Tangents is a library for working with infinite-width neural networks. It provides a high-level API for specifying complex and hierarchical neural network architectures. These networks can then be trained and evaluated either at finite-width as usual or in their infinite-width limit. Infinite-width networks can be trained analytically using exact Bayesian inference or using gradient descent via the Neural Tangent Kernel. Additionally, Neural Tangents provides tools to study gradient descent training dynamics of wide but finite networks in either function space or weight space.

The entire library runs out-of-the-box on CPU, GPU, or TPU. All computations can be automatically distributed over multiple accelerators with near-linear scaling in the number of devices.

In addition to the repository below, we provide an accompanying interactive Colab notebook at
https://colab.research.google.com/github/google/neural-tangents/blob/master/notebooks/neural_tangents_cookbook.ipynb

****
Sparse Weight Activation Training
Md Aamir Raihan,Tor M. Aamodt
Training convolutional neural networks (CNNs) is time consuming. Prior work has explored how to reduce the computational demands of training by eliminating gradients with relatively small magnitude. We show that eliminating small magnitude components has limited impact on the direction of high-dimensional vectors. However, in the context of training a CNN, we find that eliminating small magnitude components of weight and activation vectors allows us to train deeper networks on more complex datasets versus eliminating small magnitude components of gradients. We propose Sparse Weight Activation Training (SWAT), an algorithm that embodies these observations. SWAT reduces computations by 50% to 80% with better accuracy at a given level of sparsity versus the Dynamic Sparse Graph algorithm. SWAT also reduces memory footprint by 23% to 37% for activations and 50% to 80% for weights.
****
Learning Robust Representations via Multi-View Information Bottleneck
Marco Federici,Anjan Dutta,Patrick Forré,Nate Kushman,Zeynep Akata
The information bottleneck principle provides an information-theoretic method for representation learning, by training an encoder to retain all information which is relevant for predicting the label while minimizing the amount of other, excess information in the representation. The original formulation, however, requires labeled data to identify the superfluous information. In this work, we extend this ability to the multi-view unsupervised setting, where two views of the same underlying entity are provided but the label is unknown. This enables us to identify superfluous information as that not shared by both views. A theoretical analysis leads to the definition of a new multi-view model that produces state-of-the-art results on the Sketchy dataset and label-limited versions of the MIR-Flickr dataset. We also extend our theory to the single-view setting by taking advantage of standard data augmentation techniques, empirically showing better generalization capabilities when compared to common unsupervised approaches for representation learning.
****
Batch-shaping for learning conditional channel gated networks
Babak Ehteshami Bejnordi,Tijmen Blankevoort,Max Welling
We present a method that trains large capacity neural networks with significantly improved accuracy and lower dynamic computational cost. This is achieved by gating the deep-learning architecture on a fine-grained-level. Individual convolutional maps are turned on/off conditionally on features in the network. To achiev

e this, we introduce a new residual block architecture that gates convolutional channels in a fine-grained manner. We also introduce a generally applicable tool batch-shaping that matches the marginal aggregate posteriors of features in a neural network to a pre-specified prior distribution. We use this novel technique to force gates to be more conditional on the data. We present results on CIFAR-10 and ImageNet datasets for image classification, and Cityscapes for semantic segmentation. Our results show that our method can slim down large architectures conditionally, such that the average computational cost on the data is on par with a smaller architecture, but with higher accuracy. In particular, on ImageNet, our ResNet50 and ResNet34 gated networks obtain 74.60% and 72.55% top-1 accuracy compared to the 69.76% accuracy of the baseline ResNet18 model, for similar complexity. We also show that the resulting networks automatically learn to use more features for difficult examples and fewer features for simple examples.
****

Making the Shoe Fit: Architectures, Initializations, and Tuning for Learning with Privacy
Nicolas Papernot,Steve Chien,Shuang Song,Abhradeep Thakurta,Ulfar Erlingsson
Because learning sometimes involves sensitive data, standard machine-learning algorithms have been extended to offer strong privacy guarantees for training data. However, in practice, this has been mostly an afterthought, with privacy-preserving models obtained by re-running training with a different optimizer, but using the same model architecture that performed well in a non-privacy-preserving setting. This approach leads to less than ideal privacy/utility tradeoffs, as we show here. Instead, we propose that model architectures and initializations are chosen and hyperparameter tuning is performed, ab initio, explicitly for privacy-preserving training. Using this paradigm, we achieve new state-of-the-art accuracy on MNIST, FashionMNIST, and CIFAR10 without any modification of the fundamental learning procedures or differential-privacy analysis.
****

Universal Adversarial Attack Using Very Few Test Examples
Amit Deshpande,Sandesh Kamath,K V Subrahmanyam
Adversarial attacks such as Gradient-based attacks, Fast Gradient Sign Method (FGSM) by Goodfellow et al.(2015) and DeepFool by Moosavi-Dezfooli et al. (2016) are input-dependent, small pixel-wise perturbations of images which fool state of the art neural networks into misclassifying images but are unlikely to fool any human. On the other hand a universal adversarial attack is an input-agnostic perturbation. The same perturbation is applied to all inputs and yet the neural network is fooled on a large fraction of the inputs. In this paper, we show that multiple known input-dependent pixel-wise perturbations share a common spectral property. Using this spectral property, we show that the top singular vector of input-dependent adversarial attack directions can be used as a very simple universal adversarial attack on neural networks. We evaluate the error rates and fooling rates of three universal attacks, SVD-Gradient, SVD-DeepFool and SVD-FGSM, on state of the art neural networks. We show that these universal attack vectors can be computed using a small sample of test inputs. We establish our results both theoretically and empirically. On VGG19 and VGG16, the fooling rate of SVD-DeepFool and SVD-Gradient perturbations constructed from observing less than 0.2% of the validation set of ImageNet is as good as the universal attack of Moosavi-Dezfooli et al. (2017a). To prove our theoretical results, we use matrix concentration inequalities and spectral perturbation bounds. For completeness, we also discuss another recent approach to universal adversarial perturbations based on (p, q)-singular vectors, proposed independently by Khrulkov & Oseledets (2018), and point out the simplicity and efficiency of our universal attack as the key difference.
****

Rotation-invariant clustering of neuronal responses in primary visual cortex
Ivan Ustyuzhaninov,Santiago A. Cadena,Emmanouil Froudarakis,Paul G. Fahey,Edgar Y. Walker,Erick Cobos,Jacob Reimer,Fabian H. Sinz,Andreas S. Tolias,Matthias Bethge,Alexander S. Ecker
Similar to a convolutional neural network (CNN), the mammalian retina encodes vi

sual information into several dozen nonlinear feature maps, each formed by one g anglion cell type that tiles the visual space in an approximately shift-equivari ant manner. Whether such organization into distinct cell types is maintained at the level of cortical image processing is an open question. Predictive models bu ilding upon convolutional features have been shown to provide state-of-the-art p erformance, and have recently been extended to include rotation equivariance in order to account for the orientation selectivity of V1 neurons. However, general ly no direct correspondence between CNN feature maps and groups of individual ne urons emerges in these models, thus rendering it an open question whether V1 neu rons form distinct functional clusters. Here we build upon the rotation-equivari ant representation of a CNN-based V1 model and propose a methodology for cluster ing the representations of neurons in this model to find functional cell types i ndependent of preferred orientations of the neurons. We apply this method to a d ataset of 6000 neurons and visualize the preferred stimuli of the resulting clus ters. Our results highlight the range of non-linear computations in mouse V1.
****

Solving single-objective tasks by preference multi-objective reinforcement learn ing

Jinsheng Ren,Shangqi Guo,Feng Chen

There ubiquitously exist many single-objective tasks in the real world that are inevitably related to some other objectives and influenced by them. We call such task as the objective-constrained task, which is inherently a multi-objective p roblem. Due to the conflict among different objectives, a trade-off is needed. A common compromise is to design a scalar reward function through clarifying the relationship among these objectives using the prior knowledge of experts. Howeve r, reward engineering is extremely cumbersome. This will result in behaviors tha t optimize our reward function without actually satisfying our preferences. In t his paper, we explicitly cast the objective-constrained task as preference multi -objective reinforcement learning, with the overall goal of finding a Pareto opt imal policy. Combined with Trajectory Preference Domination we propose, a weight vector that reflects the agent's preference for each objective can be learned. We analyzed the feasibility of our algorithm in theory, and further proved in ex periments its better performance compared to those that design the reward functi on by experts.
****

Deep automodulators

Ari Heljakka,Yuxin Hou,Juho Kannala,Arno Solin

We introduce a novel autoencoder model that deviates from traditional autoencode rs by using the full latent vector to independently modulate each layer in the d ecoder. We demonstrate how such an 'automodulator' allows for a principled appro ach to enforce latent space disentanglement, mixing of latent codes, and a strai ghtforward way to utilize prior information that can be construed as a scale-spe cific invariance. Unlike GANs, autoencoder models can directly operate on new re al input samples. This makes our model directly suitable for applications involv ing real-world inputs. As the architectural backbone, we extend recent generativ e autoencoder models that retain input identity and image sharpness at high reso lutions better than VAEs. We show that our model achieves state-of-the-art laten t space disentanglement and achieves high quality and diversity of output sample s, as well as faithfulness of reconstructions.
****

Enhanced Convolutional Neural Tangent Kernels

Dingli Yu,Ruosong Wang,Zhiyuan Li,Wei Hu,Ruslan Salakhutdinov,Sanjeev Arora,Simo n S. Du

Recent research shows that for training with l2 loss, convolutional neural netwo rks (CNNs) whose width (number of channels in convolutional layers) goes to infi nity, correspond to regression with respect to the CNN Gaussian Process kernel (CNN-GP) if only the last layer is trained, and correspond to regression with res pect to the Convolutional Neural Tangent Kernel (CNTK) if all layers are trained . An exact algorithm to compute CNTK (Arora et al., 2019) yielded the finding th at classification accuracy of CNTK on CIFAR-10 is within 6-7% of that of the cor

responding CNN architecture (best figure being around 78%) which is interesting performance for a fixed kernel.
Here we show how to significantly enhance the performance of these kernels using two ideas. (1) Modifying the kernel using a new operation called Local Average Pooling (LAP) which preserves efficient computability of the kernel and inherits the spirit of standard data augmentation using pixel shifts. Earlier papers were unable to incorporate naive data augmentation because of the quadratic training cost of kernel regression. This idea is inspired by Global Average Pooling (GAP), which we show for CNN-GP and CNTK, GAP is equivalent to full translation data augmentation. (2) Representing the input image using a pre-processing technique proposed by Coates et al. (2011), which uses a single convolutional layer composed of random image patches.
On CIFAR-10 the resulting kernel, CNN-GP with LAP and horizontal flip data augmentation achieves 89% accuracy, matching the performance of AlexNet (Krizhevsky et al., 2012). Note that this is the best such result we know of for a classifier that is not a trained neural network. Similar improvements are obtained for Fashion-MNIST.
****

Revisiting Gradient Episodic Memory for Continual Learning
Zhiyi Chen,Tong Lin*
Gradient Episodic Memory (GEM) is an effective model for continual learning, where each gradient update for the current task is formulated as a quadratic program problem with inequality constraints that alleviate catastrophic forgetting of previous tasks. However, practical use of GEM is impeded by several limitations: (1) the data examples stored in the episodic memory may not be representative of past tasks; (2) the inequality constraints appear to be rather restrictive for competing or conflicting tasks; (3) the inequality constraints can only avoid catastrophic forgetting but can not assure positive backward transfer. To address these issues, in this paper we aim at improving the original GEM model via three handy techniques without extra computational cost. Experiments on MNIST Permutations and incremental CIFAR100 datasets demonstrate that our techniques enhance the performance of GEM remarkably. On CIFAR100 the average accuracy is improved from 66.48% to 68.76%, along with the backward (knowledge) transfer growing from 1.38% to 4.03%.
****

Inductive and Unsupervised Representation Learning on Graph Structured Objects
Lichen Wang,Bo Zong,Qianqian Ma,Wei Cheng,Jingchao Ni,Wenchao Yu,Yanchi Liu,Dongjin Song,Haifeng Chen,Yun Fu
Inductive and unsupervised graph learning is a critical technique for predictive or information retrieval tasks where label information is difficult to obtain. It is also challenging to make graph learning inductive and unsupervised at the same time, as learning processes guided by reconstruction error based loss functions inevitably demand graph similarity evaluation that is usually computationally intractable. In this paper, we propose a general framework SEED (Sampling, Encoding, and Embedding Distributions) for inductive and unsupervised representation learning on graph structured objects. Instead of directly dealing with the computational challenges raised by graph similarity evaluation, given an input graph, the SEED framework samples a number of subgraphs whose reconstruction errors could be efficiently evaluated, encodes the subgraph samples into a collection of subgraph vectors, and employs the embedding of the subgraph vector distribution as the output vector representation for the input graph. By theoretical analysis, we demonstrate the close connection between SEED and graph isomorphism. Using public benchmark datasets, our empirical study suggests the proposed SEED framework is able to achieve up to 10% improvement, compared with competitive baseline methods.
****

A new perspective in understanding of Adam-Type algorithms and beyond
Zeyi Tao,Qi Xia,Qun Li
First-order adaptive optimization algorithms such as Adam play an important role in modern deep learning due to their super fast convergence speed in solving la

rge scale optimization problems. However, Adam's non-convergence behavior and re grettable generalization ability make it fall into a love-hate relationship to d eep learning community. Previous studies on Adam and its variants (refer as Adam -Type algorithms) mainly rely on theoretical regret bound analysis, which overlo ok the natural characteristic reside in such algorithms and limit our thinking. In this paper, we aim at seeking a different interpretation of Adam-Type algorit hms so that we can intuitively comprehend and improve them. The way we chose is based on a traditional online convex optimization algorithm scheme known as mirr or descent method. By bridging Adam and mirror descent, we receive a clear map o f the functionality of each part in Adam. In addition, this new angle brings us a new insight on identifying the non-convergence issue of Adam. Moreover, we pro vide new variant of Adam-Type algorithm, namely AdamAL which can naturally mitig ate the non-convergence issue of Adam and improve its performance. We further co nduct experiments on various popular deep learning tasks and models, and the res ults are quite promising.
****

Causally Correct Partial Models for Reinforcement Learning
Danilo J. Rezende,Ivo Danihelka,George Papamakarios,Nan Rosemary Ke,Ray Jiang,Th eophane Weber,Karol Gregor,Hamza Merzic,Fabio Viola,Jane Wang,Jovana Mitrovic,Fr ederic Besse,Ioannis Antonoglou,Lars Buesing,Julian Schrittwieser,Thomas Hubert, David Silver
In reinforcement learning, we can learn a model of future observations and rewar ds, and use it to plan the agent's next actions. However, jointly modeling futur e observations can be computationally expensive or even intractable if the obser vations are high-dimensional (e.g. images). For this reason, previous works have considered partial models, which model only part of the observation. In this pa per, we show that partial models can be causally incorrect: they are confounded by the observations they don't model, and can therefore lead to incorrect planni ng. To address this, we introduce a general family of partial models that are pr ovably causally correct, but avoid the need to fully model future observations.
****

Spectral Nonlocal Block for Neural Network
Lei Zhu,Qi She,Lidan Zhang,Ping guo
The nonlocal network is designed for capturing long-range spatial-temporal depen dencies in several computer vision tasks. Although having shown excellent perfor mances, it needs an elaborate preparation for both the number and position of th e building blocks. In this paper, we propose a new formulation of the nonlocal b lock and interpret it from the general graph signal processing perspective, wher e we view it as a fully-connected graph filter approximated by Chebyshev polynom ials. The proposed nonlocal block is more efficient and robust, which is a gener alized form of existing nonlocal blocks (e.g. nonlocal block, nonlocal stage). M oreover, we give the stable hypothesis and show that the steady-state of the dee per nonlocal structure should meet with it. Based on the stable hypothesis, a f ull-order approximation of the nonlocal block is derived for consecutive connect ions. Experimental results illustrate the clear-cut improvement and practical ap plicability of the generalized nonlocal block on both image and video classifica tion tasks.
****

U-GAT-IT: Unsupervised Generative Attentional Networks with Adaptive Layer-Insta nce Normalization for Image-to-Image Translation
Junho Kim,Minjae Kim,Hyeonwoo Kang,Kwang Hee Lee
We propose a novel method for unsupervised image-to-image translation, which inc orporates a new attention module and a new learnable normalization function in a n end-to-end manner. The attention module guides our model to focus on more impo rtant regions distinguishing between source and target domains based on the atte ntion map obtained by the auxiliary classifier. Unlike previous attention-based method which cannot handle the geometric changes between domains, our model can translate both images requiring holistic changes and images requiring large shap e changes. Moreover, our new AdaLIN (Adaptive Layer-Instance Normalization) func tion helps our attention-guided model to flexibly control the amount of change i

n shape and texture by learned parameters depending on datasets. Experimental re
sults show the superiority of the proposed method compared to the existing state
-of-the-art models with a fixed network architecture and hyper-parameters.
****

Masked Based Unsupervised Content Transfer
Ron Mokady,Sagie Benaim,Lior Wolf,Amit Bermano
We consider the problem of translating, in an unsupervised manner, between two d
omains where one contains some additional information compared to the other. The
 proposed method disentangles the common and separate parts of these domains an
d, through the generation of a mask, focuses the attention of the underlying net
work to the desired augmentation alone, without wastefully reconstructing the en
tire target. This enables state-of-the-art quality and variety of content transl
ation, as demonstrated through extensive quantitative and qualitative evaluation
. Our method is also capable of adding the separate content of different guide i
mages and domains as well as remove existing separate content. Furthermore, our
method enables weakly-supervised semantic segmentation of the separate part of e
ach domain, where only class labels are provided. Our code is available at https
://github.com/rmokady/mbu-content-tansfer.

****

Efficient meta reinforcement learning via meta goal generation
Haotian Fu,Hongyao Tang,Jianye Hao
Meta reinforcement learning (meta-RL) is able to accelerate the acquisition of n
ew tasks by learning from past experience. Current meta-RL methods usually learn
 to adapt to new tasks by directly optimizing the parameters of policies over pr
imitive actions. However, for complex tasks which requires sophisticated control
 strategies, it would be quite inefficient to to directly learn such a meta-poli
cy. Moreover, this problem can become more severe and even fail in spare reward
settings, which is quite common in practice. To this end, we propose a new meta-
RL algorithm called meta goal-generation for hierarchical RL (MGHRL) by leveragi
ng hierarchical actor-critic framework. Instead of directly generate policies ov
er primitive actions for new tasks, MGHRL learns to generate high-level meta str
ategies over subgoals given past experience and leaves the rest of how to achiev
e subgoals as independent RL subtasks. Our empirical results on several challeng
ing simulated robotics environments show that our method enables more efficient
and effective meta-learning from past experience and outperforms state-of-the-ar
t meta-RL and Hierarchical-RL methods in sparse reward settings.
****

Learning robust visual representations using data augmentation invariance
Alex Hernandez-Garcia,Peter König,Tim C. Kietzmann
Deep convolutional neural networks trained for image object categorization have
shown remarkable similarities with representations found across the primate vent
ral visual stream. Yet, artificial and biological networks still exhibit importa
nt differences. Here we investigate one such property: increasing invariance to
identity-preserving image transformations found along the ventral stream. Despit
e theoretical evidence that invariance should emerge naturally from the optimiza
tion process, we present empirical evidence that the activations of convolutiona
l neural networks trained for object categorization are not robust to identity-p
reserving image transformations commonly used in data augmentation. As a solutio
n, we propose data augmentation invariance, an unsupervised learning objective w
hich improves the robustness of the learned representations by promoting the sim
ilarity between the activations of augmented image samples. Our results show tha
t this approach is a simple, yet effective and efficient (10 % increase in train
ing time) way of increasing the invariance of the models while obtaining similar
 categorization performance.
****

A Simple Dynamic Learning Rate Tuning Algorithm For Automated Training of DNNs
Koyel Mukherjee,Alind Khare,Yogish Sabharwal,Ashish Verma
Training neural networks on image datasets generally require extensive experimen
tation to find the optimal learning rate regime. Especially, for the cases of ad

versarial training or for training a newly synthesized model, one would not know the best learning rate regime beforehand. We propose an automated algorithm for determining the learning rate trajectory, that works across datasets and models for both natural and adversarial training, without requiring any dataset/model specific tuning. It is a stand-alone, parameterless, adaptive approach with no computational overhead. We theoretically discuss the algorithm's convergence behavior. We empirically validate our algorithm extensively. Our results show that our proposed approach \emph{consistently} achieves top-level accuracy compared to SOTA baselines in the literature in natural training, as well as in adversarial training.
****

DropEdge: Towards Deep Graph Convolutional Networks on Node Classification
Yu Rong,Wenbing Huang,Tingyang Xu,Junzhou Huang
Over-fitting and over-smoothing are two main obstacles of developing deep Graph Convolutional Networks (GCNs) for node classification. In particular, over-fitting weakens the generalization ability on small dataset, while over-smoothing impedes model training by isolating output representations from the input features with the increase in network depth. This paper proposes DropEdge, a novel and flexible technique to alleviate both issues. At its core, DropEdge randomly removes a certain number of edges from the input graph at each training epoch, acting like a data augmenter and also a message passing reducer. Furthermore, we theoretically demonstrate that DropEdge either reduces the convergence speed of over-smoothing or relieves the information loss caused by it. More importantly, our DropEdge is a general skill that can be equipped with many other backbone models (e.g. GCN, ResGCN, GraphSAGE, and JKNet) for enhanced performance. Extensive experiments on several benchmarks verify that DropEdge consistently improves the performance on a variety of both shallow and deep GCNs. The effect of DropEdge on preventing over-smoothing is empirically visualized and validated as well. Codes are released on~https://github.com/DropEdge/DropEdge.
****

Projected Canonical Decomposition for Knowledge Base Completion
Timothée Lacroix,Guillaume Obozinski,Joan Bruna,Nicolas Usunier
The leading approaches to tensor completion and link prediction are based on the canonical polyadic (CP) decomposition of tensors. While these approaches were originally motivated by low rank approximations, the best performances are usually obtained for ranks as high as permitted by computation constraints. For large scale factorization problems where the factor dimensions have to be kept small, the performances of these approaches tend to drop drastically. The other main tensor factorization model, Tucker decomposition, is more flexible than CP for fixed factor dimensions, so we expect Tucker-based approaches to yield better performance under strong constraints on the number of parameters. However, as we show in this paper through experiments on standard benchmarks of link prediction in knowledge bases, ComplEx, a variant of CP, achieves similar performances to recent approaches based on Tucker decomposition on all operating points in terms of number of parameters. In a control experiment, we show that one problem in the practical application of Tucker decomposition to large-scale tensor completion comes from the adaptive optimization algorithms based on diagonal rescaling, such as Adagrad. We present a new algorithm for a constrained version of Tucker which implicitly applies Adagrad to a CP-based model with an additional projection of the embeddings onto a fixed lower dimensional subspace. The resulting Tucker-style extension of ComplEx obtains similar best performances as ComplEx, with substantial gains on some datasets under constraints on the number of parameters.
****

Sequential Latent Knowledge Selection for Knowledge-Grounded Dialogue
Byeongchang Kim,Jaewoo Ahn,Gunhee Kim
Knowledge-grounded dialogue is a task of generating an informative response based on both discourse context and external knowledge. As we focus on better modeling the knowledge selection in the multi-turn knowledge-grounded dialogue, we propose a sequential latent variable model as the first approach to this matter. The model named sequential knowledge transformer (SKT) can keep track of the prior

and posterior distribution over knowledge; as a result, it can not only reduce the ambiguity caused from the diversity in knowledge selection of conversation but also better leverage the response information for proper choice of knowledge. Our experimental results show that the proposed model improves the knowledge selection accuracy and subsequently the performance of utterance generation. We achieve the new state-of-the-art performance on Wizard of Wikipedia (Dinan et al., 2019) as one of the most large-scale and challenging benchmarks. We further validate the effectiveness of our model over existing conversation methods in another knowledge-based dialogue Holl-E dataset (Moghe et al., 2018).
****
Measuring the Reliability of Reinforcement Learning Algorithms
Stephanie C.Y. Chan,Samuel Fishman,Anoop Korattikara,John Canny,Sergio Guadarrama
Lack of reliability is a well-known issue for reinforcement learning (RL) algorithms. This problem has gained increasing attention in recent years, and efforts to improve it have grown substantially. To aid RL researchers and production users with the evaluation and improvement of reliability, we propose a set of metrics that quantitatively measure different aspects of reliability. In this work, we focus on variability and risk, both during training and after learning (on a fixed policy). We designed these metrics to be general-purpose, and we also designed complementary statistical tests to enable rigorous comparisons on these metrics. In this paper, we first describe the desired properties of the metrics and their design, the aspects of reliability that they measure, and their applicability to different scenarios. We then describe the statistical tests and make additional practical recommendations for reporting results. The metrics and accompanying statistical tools have been made available as an open-source library. We apply our metrics to a set of common RL algorithms and environments, compare them, and analyze the results.
****
Stable Rank Normalization for Improved Generalization in Neural Networks and GANs
Amartya Sanyal,Philip H. Torr,Puneet K. Dokania
Exciting new work on generalization bounds for neural networks (NN) given by Bartlett et al. (2017); Neyshabur et al. (2018) closely depend on two parameter- dependant quantities: the Lipschitz constant upper bound and the stable rank (a softer version of rank). Even though these bounds typically have minimal practical utility, they facilitate questions on whether controlling such quantities together could improve the generalization behaviour of NNs in practice. To this end, we propose stable rank normalization (SRN), a novel, provably optimal, and computationally efficient weight-normalization scheme which minimizes the stable rank of a linear operator. Surprisingly we find that SRN, despite being non-convex, can be shown to have a unique optimal solution. We provide extensive analyses across a wide variety of NNs (DenseNet, WideResNet, ResNet, Alexnet, VGG), where applying SRN to their linear layers leads to improved classification accuracy, while simultaneously showing improvements in genealization, evaluated empirically using—(a) shattering experiments (Zhang et al., 2016); and (b) three measures of sample complexity by Bartlett et al. (2017), Neyshabur et al. (2018), & Wei & Ma. Additionally, we show that, when applied to the discriminator of GANs, it improves Inception, FID, and Neural divergence scores, while learning mappings with low empirical Lipschitz constant.
****
Graph Neural Networks for Soft Semi-Supervised Learning on Hypergraphs
Naganand Yadati,Tingran Gao,Shahab Asoodeh,Partha Talukdar,Anand Louis
Graph-based semi-supervised learning (SSL) assigns labels to initially unlabelled vertices in a graph.
Graph neural networks (GNNs), esp. graph convolutional networks (GCNs), inspired the current-state-of-the art models for graph-based SSL problems.
GCNs inherently assume that the labels of interest are numerical or categorical variables.
However, in many real-world applications such as co-authorship networks, recomme

ndation networks, etc., vertex labels can be naturally represented by probabilit
y distributions or histograms.
Moreover, real-world network datasets have complex relationships going beyond pa
irwise associations.
These relationships can be modelled naturally and flexibly by hypergraphs.
In this paper, we explore GNNs for graph-based SSL of histograms.
Motivated by complex relationships (those going beyond pairwise) in real-world n
etworks, we propose a novel method for directed hypergraphs.
Our work builds upon existing works on graph-based SSL of histograms derived fro
m the theory of optimal transportation.
A key contribution of this paper is to establish generalisation error bounds for
 a one-layer GNN within the framework of algorithmic stability.
We also demonstrate our proposed methods' effectiveness through detailed experim
entation on real-world data.
We have made the code available.
****

Why Not to Use Zero Imputation? Correcting Sparsity Bias in Training Neural Netw
orks

Joonyoung Yi,Juhyuk Lee,Kwang Joon Kim,Sung Ju Hwang,Eunho Yang

Handling missing data is one of the most fundamental problems in machine learnin
g. Among many approaches, the simplest and most intuitive way is zero imputation
, which treats the value of a missing entry simply as zero. However, many studie
s have experimentally confirmed that zero imputation results in suboptimal perfo
rmances in training neural networks. Yet, none of the existing work has explaine
d what brings such performance degradations. In this paper, we introduce the var
iable sparsity problem (VSP), which describes a phenomenon where the output of a
 predictive model largely varies with respect to the rate of missingness in the
given input, and show that it adversarially affects the model performance. We fi
rst theoretically analyze this phenomenon and propose a simple yet effective tec
hnique to handle missingness, which we refer to as Sparsity Normalization (SN),
that directly targets and resolves the VSP. We further experimentally validate S
N on diverse benchmark datasets, to show that debiasing the effect of input-leve
l sparsity improves the performance and stabilizes the training of neural networ
ks.
****

Self-Imitation Learning via Trajectory-Conditioned Policy for Hard-Exploration T
asks

Yijie Guo,Jongwook Choi,Marcin Moczulski,Samy Bengio,Mohammad Norouzi,Honglak Le
e

Imitation learning from human-expert demonstrations has been shown to be greatly
 helpful for challenging reinforcement learning problems with sparse environment
 rewards. However, it is very difficult to achieve similar success without relyi
ng on expert demonstrations. Recent works on self-imitation learning showed that
 imitating the agent's own past good experience could indirectly drive explorati
on in some environments, but these methods often lead to sub-optimal and myopic
behavior. To address this issue, we argue that exploration in diverse directions
 by imitating diverse trajectories, instead of focusing on limited good trajecto
ries, is more desirable for the hard-exploration tasks. We propose a new method
of learning a trajectory-conditioned policy to imitate diverse trajectories from
 the agent's own past experiences and show that such self-imitation helps avoid
myopic behavior and increases the chance of finding a globally optimal solution
for hard-exploration tasks, especially when there are misleading rewards. Our me
thod significantly outperforms existing self-imitation learning and count-based
exploration methods on various hard-exploration tasks with local optima. In part
icular, we report a state-of-the-art score of more than 20,000 points on Montezu
mas Revenge without using expert demonstrations or resetting to arbitrary states
.
****

ICNN: INPUT-CONDITIONED FEATURE REPRESENTATION LEARNING FOR TRANSFORMATION-INVAR
IANT NEURAL NETWORK

Suraj Tripathi,Chirag Singh,Abhay Kumar
We propose a novel framework, ICNN, which combines the input-conditioned filter generation module and a decoder based network to incorporate contextual information present in images into Convolutional Neural Networks (CNNs). In contrast to traditional CNNs, we do not employ the same set of learned convolution filters for all input image instances. And our proposed decoder network serves the purpose of reducing the transformation present in the input image by learning to construct a representative image of the input image class. Our proposed joint supervision of input-aware framework when combined with techniques inspired by Multi-instance learning and max-pooling, results in a transformation-invariant neural network. We investigated the performance of our proposed framework on three MNIST variations, which covers both rotation and scaling variance, and achieved 0.98% error on MNIST-rot-12k, 1.12% error on Half-rotated MNIST and 0.68% error on Scaling MNIST, which is significantly better than the state-of-the-art results. Our proposed model also showcased consistent improvement on the CIFAR dataset. We make use of visualization to further prove the effectiveness of our input-aware convolution filters. Our proposed convolution filter generation framework can also serve as a plugin for any CNN based architecture and enhance its modeling capacity.
****
Data Augmentation in Training CNNs: Injecting Noise to Images
Murtaza Eren Akbiyik
Noise injection is a fundamental tool for data augmentation, and yet there is no widely accepted procedure to incorporate it with learning frameworks. This study analyzes the effects of adding or applying different noise models of varying magnitudes to Convolutional Neural Network (CNN) architectures. Noise models that are distributed with different density functions are given common magnitude levels via Structural Similarity (SSIM) metric in order to create an appropriate ground for comparison. The basic results are conforming with the most of the common notions in machine learning, and also introduces some novel heuristics and recommendations on noise injection. The new approaches will provide better understanding on optimal learning procedures for image classification.
****
VAENAS: Sampling Matters in Neural Architecture Search
Shizheng Qin,Yichen Zhu,Pengfei Hou,Xiangyu Zhang,Wenqiang Zhang,Jian Sun
Neural Architecture Search (NAS) aims at automatically finding neural network architectures within an enormous designed search space. The search space usually contains billions of network architectures which causes extremely expensive computing costs in searching for the best-performing architecture. One-shot and gradient-based NAS approaches have recently shown to achieve superior results on various computer vision tasks such as image recognition. With the weight sharing mechanism, these methods lead to efficient model search. Despite their success, however, current sampling methods are either fixed or hand-crafted and thus ineffective. In this paper, we propose a learnable sampling module based on variational auto-encoder (VAE) for neural architecture search (NAS), named as VAENAS, which can be easily embedded into existing weight sharing NAS framework, e.g., one-shot approach and gradient-based approach, and significantly improve the performance of searching results. VAENAS generates a series of competitive results on CIFAR-10 and ImageNet in NasNet-like search space. Moreover, combined with one-shot approach, our method achieves a new state-of-the-art result for ImageNet classification model under 400M FLOPs with 77.4\% in ShuffleNet-like search space. Finally, we conduct a thorough analysis of VAENAS on NAS-bench-101 dataset, which demonstrates the effectiveness of our proposed methods.

****
Self-Educated Language Agent with Hindsight Experience Replay for Instruction Following
Geoffrey Cideron,Mathieu Seurin,Florian Strub,Olivier Pietquin
Language creates a compact representation of the world and allows the description of unlimited situations and objectives through compositionality. These propert

ies make it a natural fit to guide the training of interactive agents as it coul
d ease recurrent challenges in Reinforcement Learning such as sample complexity,
 generalization, or multi-tasking. Yet, it remains an open-problem to relate lan
guage and RL in even simple instruction following scenarios. Current methods rel
y on expert demonstrations, auxiliary losses, or inductive biases in neural arch
itectures. In this paper, we propose an orthogonal approach called Textual Hinds
ight Experience Replay (THER) that extends the Hindsight Experience Replay appro
ach to the language setting. Whenever the agent does not fulfill its instruction
, THER learn to output a new directive that matches the agent trajectory, and it
 relabels the episode with a positive reward. To do so, THER learns to map a sta
te into an instruction by using past successful trajectories, which removes the
need to have external expert interventions to relabel episodes as in vanilla HER
. We observe that this simple idea also initiates a learning synergy between lan
guage acquisition and policy learning on instruction following tasks in the Baby
AI environment.
****

Model-Agnostic Feature Selection with Additional Mutual Information
Mukund Sudarshan,Aahlad Manas Puli,Lakshmi Subramanian,Sriram Sankararaman,Rajes
h Ranganath
Answering questions about data can require understanding what parts of an input
X influence the response Y. Finding such an understanding can be built by testin
g relationships between variables through a machine learning model. For example,
 conditional randomization tests help determine whether a variable relates to th
e response given the rest of the variables. However, randomization tests require
 users to specify test statistics. We formalize a class of proper test statistic
s that are guaranteed to select a feature when it provides information about the
 response even when the rest of the features are known. We show that f-divergenc
es provide a broad class of proper test statistics. In the class of f-divergence
s, the KL-divergence yields an easy-to-compute proper test statistic that relate
s to the AMI. Questions of feature importance can be asked at the level of an in
dividual sample. We show that estimators from the same AMI test can also be use
d to find important features in a particular instance. We provide an example to
show that perfect predictive models are insufficient for instance-wise feature s
election. We evaluate our method on several simulation experiments, on a genomic
 dataset, a clinical dataset for hospital readmission, and on a subset of classe
s in ImageNet. Our method outperforms several baselines in various simulated dat
asets, is able to identify biologically significant genes, can select the most i
mportant predictors of a hospital readmission event, and is able to identify dis
tinguishing features in an image-classification task.
****

Do Deep Neural Networks for Segmentation Understand Insideness?
Kimberly M Villalobos,Vilim Stih,Amineh Ahmadinejad,Jamell Dozier,Andrew Francl,
Frederico Azevedo,Tomotake Sasaki,Xavier Boix
Image segmentation aims at grouping pixels that belong to the same object or reg
ion. At the heart of image segmentation lies the problem of determining whether
a pixel is inside or outside a region, which we denote as the "insideness" probl
em. Many Deep Neural Networks (DNNs) variants excel in segmentation benchmarks,
but regarding insideness, they have not been well visualized or understood: What
 representations do DNNs use to address the long-range relationships of insidene
ss? How do architectural choices affect the learning of these representations? I
n this paper, we take the reductionist approach by analyzing DNNs solving the in
sideness problem in isolation, i.e. determining the inside of closed (Jordan) cu
rves. We demonstrate analytically that state-of-the-art feed-forward and recurre
nt architectures can implement solutions of the insideness problem for any given
 curve. Yet, only recurrent networks could learn these general solutions when t
he training enforced a specific "routine" capable of breaking down the long-rang
e relationships. Our results highlights the need for new training strategies tha
t decompose the learning into appropriate stages, and that lead to the general c
lass of solutions necessary for DNNs to understand insideness.
****

Adversarial Robustness as a Prior for Learned Representations
Logan Engstrom,Andrew Ilyas,Shibani Santurkar,Dimitris Tsipras,Brandon Tran,Alek sander Madry
An important goal in deep learning is to learn versatile, high-level feature rep resentations of input data. However, standard networks' representations seem to possess shortcomings that, as we illustrate, prevent them from fully realizing t his goal. In this work, we show that robust optimization can be re-cast as a too l for enforcing priors on the features learned by deep neural networks. It turns out that representations learned by robust models address the aforementioned sh ortcomings and make significant progress towards learning a high-level encoding of inputs. In particular, these representations are approximately invertible, wh ile allowing for direct visualization and manipulation of salient input features . More broadly, our results indicate adversarial robustness as a promising avenu e for improving learned representations.
****
Explaining Time Series by Counterfactuals
Sana Tonekaboni,Shalmali Joshi,David Duvenaud,Anna Goldenberg
We propose a method to automatically compute the importance of features at every observation in time series, by simulating counterfactual trajectories given pre vious observations. We define the importance of each observation as the change i n the model output caused by replacing the observation with a generated one. Our method can be applied to arbitrarily complex time series models. We compare the generated feature importance to existing methods like sensitivity analyses, fea ture occlusion, and other explanation baselines to show that our approach genera tes more precise explanations and is less sensitive to noise in the input signal s.
****
Variational Diffusion Autoencoders with Random Walk Sampling
Henry Li,Ofir Lindenbaum,Xiuyuan Cheng,Alexander Cloninger
Variational inference (VI) methods and especially variational autoencoders (VAEs) specify scalable generative models that enjoy an intuitive connection to manif old learning --- with many default priors the posterior/likelihood pair $q(z|x)$ /$p(x|z)$ can be viewed as an approximate homeomorphism (and its inverse) betwee n the data manifold and a latent Euclidean space. However, these approximations are well-documented to become degenerate in training. Unless the subjective prio r is carefully chosen, the topologies of the prior and data distributions often will not match.
Conversely, diffusion maps (DM) automatically \textit{infer} the data topology a nd enjoy a rigorous connection to manifold learning, but do not scale easily or provide the inverse homeomorphism.
In this paper, we propose \textbf{a)} a principled measure for recognizing the m ismatch between data and latent distributions and \textbf{b)} a method that comb ines the advantages of variational inference and diffusion maps to learn a homeo morphic generative model. The measure, the \textit{locally bi-Lipschitz property }, is a sufficient condition for a homeomorphism and easy to compute and interpr et. The method, the \textit{variational diffusion autoencoder} (VDAE), is a nove l generative algorithm that first infers the topology of the data distribution, then models a diffusion random walk over the data. To achieve efficient computat ion in VDAEs, we use stochastic versions of both variational inference and manif old learning optimization. We prove approximation theoretic results for the dime nsion dependence of VDAEs, and that locally isotropic sampling in the latent spa ce results in a random walk over the reconstructed manifold.
Finally, we demonstrate the utility of our method on various real and synthetic datasets, and show that it exhibits performance superior to other generative mod els.
****
Probability Calibration for Knowledge Graph Embedding Models
Pedro Tabacof,Luca Costabello
Knowledge graph embedding research has overlooked the problem of probability cal ibration. We show popular embedding models are indeed uncalibrated. That means p

robability estimates associated to predicted triples are unreliable. We present a novel method to calibrate a model when ground truth negatives are not availabl e, which is the usual case in knowledge graphs. We propose to use Platt scaling and isotonic regression alongside our method. Experiments on three datasets with ground truth negatives show our contribution leads to well calibrated models wh en compared to the gold standard of using negatives. We get significantly better results than the uncalibrated models from all calibration methods. We show isot onic regression offers the best the performance overall, not without trade-offs. We also show that calibrated models reach state-of-the-art accuracy without the need to define relation-specific decision thresholds.
****

Contrastive Multiview Coding
Yonglong Tian,Dilip Krishnan,Phillip Isola
Humans view the world through many sensory channels, e.g., the long-wavelength l ight channel, viewed by the left eye, or the high-frequency vibrations channel, heard by the right ear. Each view is noisy and incomplete, but important factors , such as physics, geometry, and semantics, tend to be shared between all views (e.g., a "dog" can be seen, heard, and felt). We hypothesize that a powerful rep resentation is one that models view-invariant factors. Based on this hypothesis, we investigate a contrastive coding scheme, in which a representation is learne d that aims to maximize mutual information between different views but is otherw ise compact. Our approach scales to any number of views, and is view-agnostic. T he resulting learned representations perform above the state of the art for down stream tasks such as object classification, compared to formulations based on pr edictive learning or single view reconstruction, and improve as more views are a dded. On the Imagenet linear readoff benchmark, we achieve 68.4% top-1 accuracy.

****

Reformer: The Efficient Transformer
Nikita Kitaev,Lukasz Kaiser,Anselm Levskaya
Large Transformer models routinely achieve state-of-the-art results on
a number of tasks but training these models can be prohibitively costly,
especially on long sequences. We introduce two techniques to improve
the efficiency of Transformers. For one, we replace dot-product attention
by one that uses locality-sensitive hashing, changing its complexity
from O(L^2) to O($L \log L$), where L is the length of the sequence.
Furthermore, we use reversible residual layers instead of the standard
residuals, which allows storing activations only once in the training
process instead of N times, where N is the number of layers.
The resulting model, the Reformer, performs on par with Transformer models
while being much more memory-efficient and much faster on long sequences.
****

BasisVAE: Orthogonal Latent Space for Deep Disentangled Representation
Jin-Young Kim,Sung-Bae Cho
The variational autoencoder, one of the generative models, defines the latent sp ace for the data representation, and uses variational inference to infer the pos terior probability. Several methods have been devised to disentangle the latent space for controlling the generative model easily. However, due to the excessive constraints, the more disentangled the latent space is, the lower quality the g enerative model has. A disentangled generative model would allocate a single fea ture of the generated data to the only single latent variable. In this paper, we propose a method to decompose the latent space into basis, and reconstruct it b y linear combination of the latent bases. The proposed model called BasisVAE con sists of the encoder that extracts the features of data and estimates the coeffi cients for linear combination of the latent bases, and the decoder that reconstr ucts the data with the combined latent bases. In this method, a single latent ba sis is subject to change in a single generative factor, and relatively invariant to the changes in other factors. It maintains the performance while relaxing th e constraint for disentanglement on a basis, as we no longer need to decompose l atent space on a standard basis. Experiments on the well-known benchmark dataset

s of MNIST, 3DFaces and CelebA demonstrate the efficacy of the proposed method, compared to other state-of-the-art methods. The proposed model not only defines the latent space to be separated by the generative factors, but also shows the better quality of the generated and reconstructed images. The disentangled representation is verified with the generated images and the simple classifier trained on the output of the encoder.

****

Target-Embedding Autoencoders for Supervised Representation Learning

Daniel Jarrett,Mihaela van der Schaar

Autoencoder-based learning has emerged as a staple for disciplining representations in unsupervised and semi-supervised settings. This paper analyzes a framework for improving generalization in a purely supervised setting, where the target space is high-dimensional. We motivate and formalize the general framework of target-embedding autoencoders (TEA) for supervised prediction, learning intermediate latent representations jointly optimized to be both predictable from features as well as predictive of targets---encoding the prior that variations in targets are driven by a compact set of underlying factors. As our theoretical contribution, we provide a guarantee of generalization for linear TEAs by demonstrating uniform stability, interpreting the benefit of the auxiliary reconstruction task as a form of regularization. As our empirical contribution, we extend validation of this approach beyond existing static classification applications to multivariate sequence forecasting, verifying their advantage on both linear and nonlinear recurrent architectures---thereby underscoring the further generality of this framework beyond feedforward instantiations.

****

Watch the Unobserved: A Simple Approach to Parallelizing Monte Carlo Tree Search

Anji Liu,Jianshu Chen,Mingze Yu,Yu Zhai,Xuewen Zhou,Ji Liu

Monte Carlo Tree Search (MCTS) algorithms have achieved great success on many challenging benchmarks (e.g., Computer Go). However, they generally require a large number of rollouts, making their applications costly. Furthermore, it is also extremely challenging to parallelize MCTS due to its inherent sequential nature: each rollout heavily relies on the statistics (e.g., node visitation counts) estimated from previous simulations to achieve an effective exploration-exploitation tradeoff. In spite of these difficulties, we develop an algorithm, WU-UCT, to effectively parallelize MCTS, which achieves linear speedup and exhibits only limited performance loss with an increasing number of workers. The key idea in WU-UCT is a set of statistics that we introduce to track the number of on-going yet incomplete simulation queries (named as unobserved samples). These statistics are used to modify the UCT tree policy in the selection steps in a principled manner to retain effective exploration-exploitation tradeoff when we parallelize the most time-consuming expansion and simulation steps. Experiments on a proprietary benchmark and the Atari Game benchmark demonstrate the linear speedup and the superior performance of WU-UCT comparing to existing techniques.

****

Conditional Flow Variational Autoencoders for Structured Sequence Prediction

Apratim Bhattacharyya,Michael Hanselmann,Mario Fritz,Bernt Schiele,Christoph-Nikolas Straehle

Prediction of future states of the environment and interacting agents is a key competence required for autonomous agents to operate successfully in the real world. Prior work for structured sequence prediction based on latent variable models imposes a uni-modal standard Gaussian prior on the latent variables. This induces a strong model bias which makes it challenging to fully capture the multi-modality of the distribution of the future states. In this work, we introduce Conditional Flow Variational Autoencoders (CF-VAE) using our novel conditional normalizing flow based prior to capture complex multi-modal conditional distributions for effective structured sequence prediction. Moreover, we propose two novel regularization schemes which stabilizes training and deals with posterior collapse for stable training and better match to the data distribution. Our experiments on three multi-modal structured sequence prediction datasets -- MNIST Sequences, Stanford Drone and HighD -- show that the proposed method obtains state of art

results across different evaluation metrics.
****
High-Frequency guided Curriculum Learning for Class-specific Object Boundary Det
ection
VSR Veeravasarapu,Deepak Mittal,Abhishek Goel,Maneesh Singh
This work addresses class-specific object boundary extraction, i.e., retrieving
boundary pixels that belong to a class of objects in the given image. Although r
ecent ConvNet-based approaches demonstrate impressive results, we notice that th
ey produce several false-alarms and misdetections when used in real-world applic
ations. We hypothesize that although boundary detection is simple at some pixels
 that are rooted in identifiable high-frequency locations, other pixels pose a h
igher level of difficulties, for instance, region pixels with an appearance simi
lar to the boundaries; or boundary pixels with insignificant edge strengths. The
refore, the training process needs to account for different levels of learning c
omplexity in different regions to overcome false alarms. In this work, we devise
 a curriculum-learning-based training process for object boundary detection. Thi
s multi-stage training process first trains the network at simpler pixels (with
sufficient edge strengths) and then at harder pixels in the later stages of the
curriculum. We also propose a novel system for object boundary detection that r
elies on a fully convolutional neural network (FCN) and wavelet decomposition of
 image frequencies. This system uses high-frequency bands from the wavelet pyra
mid and augments them to conv features from different layers of FCN. Our ablatio
n studies with contourMNIST dataset, a simulated digit contours from MNIST, demo
nstrate that this explicit high-frequency augmentation helps the model to conver
ge faster. Our model trained by the proposed curriculum scheme outperforms a sta
te-of-the-art object boundary detection method by a significant margin on a chal
lenging aerial image dataset.

****
On the Equivalence between Positional Node Embeddings and Structural Graph Repre
sentations
Balasubramaniam Srinivasan,Bruno Ribeiro
This work provides the first unifying theoretical framework for node (positional
) embeddings and structural graph representations, bridging methods like matrix
factorization and graph neural networks. Using invariant theory, we show that re
lationship between structural representations and node embeddings is analogous t
o that of a distribution and its samples. We prove that all tasks that can be pe
rformed by node embeddings can also be performed by structural representations a
nd vice-versa. We also show that the concept of transductive and inductive learn
ing is unrelated to node embeddings and graph representations, clearing another
source of confusion in the literature. Finally, we introduce new practical guide
lines to generating and using node embeddings, which further augments standa
rd operating procedures used today.
****
Disagreement-Regularized Imitation Learning
Kiante Brantley,Wen Sun,Mikael Henaff
We present a simple and effective algorithm designed to address the covariate sh
ift problem in imitation learning. It operates by training an ensemble of polici
es on the expert demonstration data, and using the variance of their predictions
 as a cost which is minimized with RL together with a supervised behavioral clon
ing cost. Unlike adversarial imitation methods, it uses a fixed reward function
which is easy to optimize. We prove a regret bound for the algorithm which is li
near in the time horizon multiplied by a coefficient which we show to be low for
 certain problems in which behavioral cloning fails. We evaluate our algorithm e
mpirically across multiple pixel-based Atari environments and continuous control
 tasks, and show that it matches or significantly outperforms behavioral cloning
 and generative adversarial imitation learning.
****
Shifted Randomized Singular Value Decomposition
Ali Basirat

We extend the randomized singular value decomposition (SVD) algorithm (Halko et al., 2011) to estimate the SVD of a shifted data matrix without explicitly constructing the matrix in the memory. With no loss in the accuracy of the original algorithm, the extended algorithm provides for a more efficient way of matrix factorization. The algorithm facilitates the low-rank approximation and principal component analysis (PCA) of off-center data matrices. When applied to different types of data matrices, our experimental results confirm the advantages of the extensions made to the original algorithm.
****

PassNet: Learning pass probability surfaces from single-location labels. An architecture for visually-interpretable soccer analytics
Javier Fernández,Luke Bornn
We propose a fully convolutional network architecture that is able to estimate a full surface of pass probabilities from single-location labels derived from high frequency spatio-temporal data of professional soccer matches. The network is able to perform remarkably well from low-level inputs by learning a feature hierarchy that produces predictions at different sampling levels that are merged together to preserve both coarse and fine detail. Our approach presents an extreme case of weakly supervised learning where there is just a single pixel correspondence between ground-truth outcomes and the predicted probability map. By providing not just an accurate evaluation of observed events but also a visual interpretation of the results of other potential actions, our approach opens the door for spatio-temporal decision-making analysis, an as-yet little-explored area in sports. Our proposed deep learning architecture can be easily adapted to solve many other related problems in sports analytics; we demonstrate this by extending the network to learn to estimate pass-selection likelihood.
****

On Incorporating Semantic Prior Knowlegde in Deep Learning Through Embedding-Space Constraints
Damien Teney,Ehsan Abbasnejad,Anton van den Hengel
The knowledge that humans hold about a problem often extends far beyond a set of training data and output labels. While the success of deep learning mostly relies on supervised training, important properties cannot be inferred efficiently from end-to-end annotations alone, for example causal relations or domain-specific invariances. We present a general technique to supplement supervised training with prior knowledge expressed as relations between training instances. We illustrate the method on the task of visual question answering to exploit various auxiliary annotations, including relations of equivalence and of logical entailment between questions. Existing methods to use these annotations, including auxiliary losses and data augmentation, cannot guarantee the strict inclusion of these relations into the model since they require a careful balancing against the end-to-end objective. Our method uses these relations to shape the embedding space of the model, and treats them as strict constraints on its learned representations. %The resulting model encodes relations that better generalize across instances. In the context of VQA, this approach brings significant improvements in accuracy and robustness, in particular over the common practice of incorporating the constraints as a soft regularizer. We also show that incorporating this type of prior knowledge with our method brings consistent improvements, independently from the amount of supervised data used. It demonstrates the value of an additional training signal that is otherwise difficult to extract from end-to-end annotations alone.
****

Are Few-shot Learning Benchmarks Too Simple ?
Gabriel Huang,Hugo Larochelle,Simon Lacoste-Julien
We argue that the widely used Omniglot and miniImageNet benchmarks are too simple because their class semantics do not vary across episodes, which defeats their intended purpose of evaluating few-shot classification methods. The class semantics of Omniglot is invariably “characters” and the class semantics of miniImageNet, “object category”. Because the class semantics are so similar, we propose a new method called Centroid Networks which can achieve surprisingly high accurac

ies on Omniglot and miniImageNet without using any labels at metaevaluation time . Our results suggest that those benchmarks are not adapted for supervised few-s hot classification since the supervision itself is not necessary during meta-eva luation. The Meta-Dataset, a collection of 10 datasets, was recently proposed as a harder few-shot classification benchmark. Using our method, we derive a new m etric, the Class Semantics Consistency Criterion, and use it to quantify the dif ficulty of Meta-Dataset. Finally, under some restrictive assumptions, we show th at Centroid Networks is faster and more accurate than a state-of-the-art learnin g-to-cluster method (Hsu et al., 2018).
****
UNIVERSAL MODAL EMBEDDING OF DYNAMICS IN VIDEOS AND ITS APPLICATIONS
Israr Ul Haq,Yoshinobu Kawahara
Extracting underlying dynamics of objects in image sequences is one of the chall enging problems in computer vision. On the other hand, dynamic mode decompositio n (DMD) has recently attracted attention as a way of obtaining modal representat ions of nonlinear dynamics from (general multivariate time-series) data without explicit prior knowledge about the dynamics. In this paper, we propose a convolu tional autoencoder based DMD (CAE-DMD) that is an extended DMD (EDMD) approach, to extract underlying dynamics in videos. To this end, we develop a modified CAE model by incorporating DMD on the encoder, which gives a more meaningful compre ssed representation of input image sequences. On the reconstruction side, a deco der is used to minimize the reconstruction error after applying the DMD, which i n result gives an accurate reconstruction of inputs. We empirically investigated the performance of CAE-DMD in two applications: background/foreground extractio n and video classification, on publicly available datasets.
****
Function Feature Learning of Neural Networks
Guangcong Wang,Jianhuang Lai,Guangrun Wang,Wenqi Liang
We present a Function Feature Learning (FFL) method that can measure the similar ity of non-convex neural networks. The function feature representation provides crucial insights into the understanding of the relations between different local solutions of identical neural networks. Unlike existing methods that use neuron activation vectors over a given dataset as neural network representation, FFL a ligns weights of neural networks and projects them into a common function featur e space by introducing a chain alignment rule. We investigate the function featu re representation on Multi-Layer Perceptron (MLP), Convolutional Neural Network (CNN), and Recurrent Neural Network (RNN), finding that identical neural network s trained with different random initializations on different learning tasks by t he Stochastic Gradient Descent (SGD) algorithm can be projected into different f ixed points. This finding demonstrates the strong connection between different l ocal solutions of identical neural networks and the equivalence of projected loc al solutions. With FFL, we also find that the semantics are often presented in a bottom-up way. Besides, FFL provides more insights into the structure of local solutions. Experiments on CIFAR-100, NameData, and tiny ImageNet datasets valida te the effectiveness of the proposed method.
****
Manifold Learning and Alignment with Generative Adversarial Networks
Jiseob Kim,Seungjae Jung,Hyundo Lee,Byoung-Tak Zhang
We present a generative adversarial network (GAN) that conducts manifold learnin g and alignment (MLA): A task to learn the multi-manifold structure underlying d ata and to align those manifolds without any correspondence information. Our mai n idea is to exploit the powerful abstraction ability of encoder architecture. S pecifically, we define multiple generators to model multiple manifolds, but in a particular way that their inverse maps can be commonly represented by a single smooth encoder. Then, the abstraction ability of the encoder enforces semantic s imilarities between the generators and gives a plausibly aligned embedding in th e latent space. In experiments with MNIST, 3D-Chair, and UT-Zap50k datasets, we demonstrate the superiority of our model in learning the manifolds by FID scores and in aligning the manifolds by disentanglement scores. Furthermore, by virtue of the abstractive modeling, we show that our model can generate data from an u

ntrained manifold, which is unique to our model.
****

Learning Deep-Latent Hierarchies by Stacking Wasserstein Autoencoders
Benoit Gaujac,Ilya Feige,David Barber
Probabilistic models with hierarchical-latent-variable structures provide state-of-the-art results amongst non-autoregressive, unsupervised density-based models. However, the most common approach to training such models based on Variational Autoencoders often fails to leverage deep-latent hierarchies; successful approaches require complex inference and optimisation schemes. Optimal Transport is an alternative, non-likelihood-based framework for training generative models with appealing theoretical properties, in principle allowing easier training convergence between distributions. In this work we propose a novel approach to training models with deep-latent hierarchies based on Optimal Transport, without the need for highly bespoke models and inference networks. We show that our method enables the generative model to fully leverage its deep-latent hierarchy, and that in-so-doing, it is more effective than the original Wasserstein Autoencoder with Maximum Mean Discrepancy divergence.
****

Scalable Deep Neural Networks via Low-Rank Matrix Factorization
Atsushi Yaguchi,Taiji Suzuki,Shuhei Nitta,Yukinobu Sakata,Akiyuki Tanizawa
Compressing deep neural networks (DNNs) is important for real-world applications operating on resource-constrained devices. However, it is difficult to change the model size once the training is completed, which needs re-training to configure models suitable for different devices. In this paper, we propose a novel method that enables DNNs to flexibly change their size after training. We factorize the weight matrices of the DNNs via singular value decomposition (SVD) and change their ranks according to the target size. In contrast with existing methods, we introduce simple criteria that characterize the importance of each basis and layer, which enables to effectively compress the error and complexity of models as little as possible. In experiments on multiple image-classification tasks, our method exhibits favorable performance compared with other methods.
****

NoiGAN: NOISE AWARE KNOWLEDGE GRAPH EMBEDDING WITH GAN
Kewei Cheng,Yikai Zhu,Ming Zhang,Yizhou Sun
Knowledge graph has gained increasing attention in recent years for its successful applications of numerous tasks. Despite the rapid growth of knowledge construction, knowledge graphs still suffer from severe incompletion and inevitably involve various kinds of errors. Several attempts have been made to complete knowledge graph as well as to detect noise. However, none of them considers unifying these two tasks even though they are inter-dependent and can mutually boost the performance of each other. In this paper, we proposed to jointly combine these two tasks with a unified Generative Adversarial Networks (GAN) framework to learn noise-aware knowledge graph embedding. Extensive experiments have demonstrated that our approach is superior to existing state-of-the-art algorithms both in regard to knowledge graph completion and error detection.
****

Fast Task Adaptation for Few-Shot Learning
Yingying Zhang,Qiaoyong Zhong,Di Xie,Shiliang Pu
Few-shot classification is a challenging task due to the scarcity of training examples for each class. The key lies in generalization of prior knowledge learned from large-scale base classes and fast adaptation of the classifier to novel classes. In this paper, we introduce a two-stage framework. In the first stage, we attempt to learn task-agnostic feature on base data with a novel Metric-Softmax loss. The Metric-Softmax loss is trained against the whole label set and learns more discriminative feature than episodic training. Besides, the Metric-Softmax classifier can be applied to base and novel classes in a consistent manner, which is critical for the generalizability of the learned feature. In the second stage, we design a task-adaptive transformation which adapts the classifier to each few-shot setting very fast within a few tuning epochs. Compared with existing fine-tuning scheme, the scarce examples of novel classes are exploited more effe

ctively. Experiments show that our approach outperforms current state-of-the-art s by a large margin on the commonly used mini-ImageNet and CUB-200-2011 benchmar ks.
****
Weighted Empirical Risk Minimization: Transfer Learning based on Importance Samp ling
Robin Vogel,Mastane Achab,Charles Tillier,Stéphan Clémençon
We consider statistical learning problems, when the distribution P' of the tra ining observations $Z'_1,\; \ldots,\; Z'_n$ differs from the distribution P in volved in the risk one seeks to minimize (referred to as the \textit{test distri bution}) but is still defined on the same measurable space as P and dominates it. In the unrealistic case where the likelihood ratio $\Phi(z)=dP/dP'(z)$ is kn own, one may straightforwardly extends the Empirical Risk Minimization (ERM) app roach to this specific \textit{transfer learning} setup using the same idea as t hat behind Importance Sampling, by minimizing a weighted version of the empirica l risk functional computed from the 'biased' training data Z'_i with weights $ \Phi(Z'_i)$. Although the \textit{importance function} $\Phi(z)$ is generally un known in practice, we show that, in various situations frequently encountered in practice, it takes a simple form and can be directly estimated from the Z'_i' s and some auxiliary information on the statistical population P. By means of linearization techniques, we then prove that the generalization capacity of the approach aforementioned is preserved when plugging the resulting estimates of th e $\Phi(Z'_i)$'s into the weighted empirical risk. Beyond these theoretical gua rantees, numerical results provide strong empirical evidence of the relevance of the approach promoted in this article.
****
Neural Program Synthesis By Self-Learning
Yifan Xu,Lu Dai,Udaikaran Singh,Kening Zhang,Zhuowen Tu
Neural inductive program synthesis is a task generating instructions that can pr oduce desired outputs from given inputs. In this paper, we focus on the generati on of a chunk of assembly code that can be executed to match a state change insi de the CPU. We develop a neural program synthesis algorithm, AutoAssemblet, lear ned via self-learning reinforcement learning that explores the large code space efficiently. Policy networks and value networks are learned to reduce the breadt h and depth of the Monte Carlo Tree Search, resulting in better synthesis perfor mance. We also propose an effective multi-entropy policy sampling technique to alleviate online update correlations. We apply AutoAssemblet to basic programmi ng tasks and show significant higher success rates compared to several competing baselines.
****
Neural Epitome Search for Architecture-Agnostic Network Compression
Daquan Zhou,Xiaojie Jin,Qibin Hou,Kaixin Wang,Jianchao Yang,Jiashi Feng
Traditional compression methods including network pruning, quantization, low ran k factorization and knowledge distillation all assume that network architectures and parameters should be hardwired. In this work, we propose a new perspective on network compression, i.e., network parameters can be disentangled from the a rchitectures. From this viewpoint, we present the Neural Epitome Search (NES), a new neural network compression approach that learns to find compact yet expres sive epitomes for weight parameters of a specified network architecture end-to-e nd. The complete network to compress can be generated from the learned epitome v ia a novel transformation method that adaptively transforms the epitomes to matc h shapes of the given architecture. Compared with existing compression methods, NES allows the weight tensors to be independent of the architecture design and h ence can achieve a good trade-off between model compression rate and performance given a specific model size constraint. Experiments demonstrate that, on ImageN et, when taking MobileNetV2 as backbone, our approach improves the full-model ba seline by 1.47% in top-1 accuracy with 25% MAdd reduction and AutoML for Model C ompression (AMC) by 2.5% with nearly the same compression ratio. Moreover, takin g EfficientNet-B0 as baseline, our NES yields an improvement of 1.2% but are wit h 10% less MAdd. In particular, our method achieves a new state-of-the-art resu

lts of 77.5% under mobile settings (<350M MAdd). Code will be made publicly avai
lable.
****

Learning from Label Proportions with Consistency Regularization
Kuen-Han Tsai,Hsuan-Tien Lin
The problem of learning from label proportions (LLP) involves training classifie
rs with weak labels on bags of instances, rather than strong labels on individua
l instances. The weak labels only contain the label proportions of each bag. The
 LLP problem is important for many practical applications that only allow label
proportions to be collected because of data privacy or annotation costs, and has
 recently received lots of research attention. Most existing works focus on exte
nding supervised learning models to solve the LLP problem, but the weak learning
 nature makes it hard to further improve LLP performance with a supervised angle
. In this paper, we take a different angle from semi-supervised learning.
In particular, we propose a novel model inspired by consistency regularization,
a popular concept in semi-supervised learning that encourages the model to produ
ce a decision boundary that better describes the data manifold. With the introdu
ction of consistency regularization, we further extend our study to non-uniform
bag-generation and validation-based parameter-selection procedures that better m
atch practical needs. Experiments not only justify that LLP with consistency reg
ularization achieves superior performance, but also demonstrate the practical us
ability of the proposed procedures.
****

Do recent advancements in model-based deep reinforcement learning really improve
 data efficiency?
Kacper Piotr Kielak
Reinforcement learning (RL) has seen great advancements in the past few years. N
evertheless, the consensus among the RL community is that currently used model-f
ree methods, despite all their benefits, suffer from extreme data inefficiency.
To circumvent this problem, novel model-based approaches were introduced that of
ten claim to be much more efficient than their model-free counterparts. In this
paper, however, we demonstrate that the state-of-the-art model-free Rainbow DQN
algorithm can be trained using a much smaller number of samples than it is commo
nly reported. By simply allowing the algorithm to execute network updates more f
requently we manage to reach similar or better results than existing model-based
 techniques, at a fraction of complexity and computational costs. Furthermore, b
ased on the outcomes of the study, we argue that the agent similar to the modifi
ed Rainbow DQN that is presented in this paper should be used as a baseline for
any future work aimed at improving sample efficiency of deep reinforcement learn
ing.
****

Evo-NAS: Evolutionary-Neural Hybrid Agent for Architecture Search
Krzysztof Maziarz,Mingxing Tan,Andrey Khorlin,Kuang-Yu Samuel Chang,Andrea Gesmu
ndo
Neural Architecture Search has shown potential to automate the design of neural
networks. Deep Reinforcement Learning based agents can learn complex architectur
al patterns, as well as explore a vast and compositional search space. On the ot
her hand, evolutionary algorithms offer higher sample efficiency, which is criti
cal for such a resource intensive application. In order to capture the best of b
oth worlds, we propose a class of Evolutionary-Neural hybrid agents (Evo-NAS). W
e show that the Evo-NAS agent outperforms both neural and evolutionary agents wh
en applied to architecture search for a suite of text and image classification b
enchmarks. On a high-complexity architecture search space for image classificati
on, the Evo-NAS agent surpasses the accuracy achieved by commonly used agents wi
th only 1/3 of the search cost.
****

Quaternion Equivariant Capsule Networks for 3D Point Clouds
Yongheng Zhao,Tolga Birdal,Jan Eric Lenssen,Emanuele Menegatti,Leonidas Guibas,F
ederico Tombari
We present a 3D capsule architecture for processing of point clouds that is equi

variant with respect to the SO(3) rotation group, translation and permutation of the unordered input sets. The network operates on a sparse set of local reference frames, computed from an input point cloud and establishes end-to-end equivariance through a novel 3D quaternion group capsule layer, including an equivariant dynamic routing procedure. The capsule layer enables us to disentangle geometry from pose, paving the way for more informative descriptions and a structured latent space. In the process, we theoretically connect the process of dynamic routing between capsules to the well-known Weiszfeld algorithm, a scheme for solving iterative re-weighted least squares (IRLS) problems with provable convergence properties, enabling robust pose estimation between capsule layers. Due to the sparse equivariant quaternion capsules, our architecture allows joint object classification and orientation estimation, which we validate empirically on common benchmark datasets.

****

When Covariate-shifted Data Augmentation Increases Test Error And How to Fix It

Sang Michael Xie*,Aditi Raghunathan*,Fanny Yang,John C. Duchi,Percy Liang

Empirically, data augmentation sometimes improves and sometimes hurts test error, even when only adding points with labels from the true conditional distribution that the hypothesis class is expressive enough to fit. In this paper, we provide precise conditions under which data augmentation hurts test accuracy for minimum norm estimators in linear regression. To mitigate the failure modes of augmentation, we introduce X-regularization, which uses unlabeled data to regularize the parameters towards the non-augmented estimate. We prove that our new estimator never hurts test error and exhibits significant improvements over adversarial data augmentation on CIFAR-10.

****

Accelerated Variance Reduced Stochastic Extragradient Method for Sparse Machine Learning Problems

Fanhua Shang,Lin Kong,Yuanyuan Liu,Hua Huang,Hongying Liu

Recently, many stochastic gradient descent algorithms with variance reduction have been proposed. Moreover, their proximal variants such as Prox-SVRG can effectively solve non-smooth problems, which makes that they are widely applied in many machine learning problems. However, the introduction of proximal operator will result in the error of the optimal value. In order to address this issue, we introduce the idea of extragradient and propose a novel accelerated variance reduced stochastic extragradient descent (AVR-SExtraGD) algorithm, which inherits the advantages of Prox-SVRG and momentum acceleration techniques. Moreover, our theoretical analysis shows that AVR-SExtraGD enjoys the best-known convergence rates and oracle complexities of stochastic first-order algorithms such as Katyusha for both strongly convex and non-strongly convex problems. Finally, our experimental results show that for ERM problems and robust face recognition via sparse representation, our AVR-SExtraGD can yield the improved performance compared with Prox-SVRG and Katyusha. The asynchronous variant of AVR-SExtraGD outperforms KroMagnon and ASAGA, which are the asynchronous variants of SVRG and SAGA, respectively.

****

The Variational InfoMax AutoEncoder

Vinenzo Crescimanna,Bruce Graham

We propose the Variational InfoMax AutoEncoder (VIMAE), an autoencoder based on a new learning principle for unsupervised models: the Capacity-Constrained InfoMax, which allows the learning of a disentangled representation while maintaining optimal generative performance. The variational capacity of an autoencoder is defined and we investigate its role. We associate the two main properties of a Variational AutoEncoder (VAE), generation quality and disentangled representation, to two different information concepts, respectively Mutual Information and network capacity. We deduce that a small capacity autoencoder tends to learn a more robust and disentangled representation than a high capacity one. This observation is confirmed by the computational experiments.

****
Skew-Fit: State-Covering Self-Supervised Reinforcement Learning
Vitchyr H. Pong,Murtaza Dalal,Steven Lin,Ashvin Nair,Shikhar Bahl,Sergey Levine
Autonomous agents that must exhibit flexible and broad capabilities will need to be equipped with large repertoires of skills. Defining each skill with a manually-designed reward function limits this repertoire and imposes a manual engineering burden. Self-supervised agents that set their own goals can automate this process, but designing appropriate goal setting objectives can be difficult, and often involves heuristic design decisions. In this paper, we propose a formal exploration objective for goal-reaching policies that maximizes state coverage. We show that this objective is equivalent to maximizing the entropy of the goal distribution together with goal reaching performance, where goals correspond to full state observations. To instantiate this principle, we present an algorithm called Skew-Fit for learning a maximum-entropy goal distributions. Skew-Fit enables self-supervised agents to autonomously choose and practice reaching diverse goals. We show that, under certain regularity conditions, our method converges to a uniform distribution over the set of valid states, even when we do not know this set beforehand. Our experiments show that it can learn a variety of manipulation tasks from images, including opening a door with a real robot, entirely from scratch and without any manually-designed reward function.
****
LOGAN: Latent Optimisation for Generative Adversarial Networks
Yan Wu,Jeff Donahue,David Balduzzi,Karen Simonyan,Timothy Lillicrap
Training generative adversarial networks requires balancing of delicate adversarial dynamics. Even with careful tuning, training may diverge or end up in a bad equilibrium with dropped modes. In this work, we introduce a new form of latent optimisation inspired by the CS-GAN and show that it improves adversarial dynamics by enhancing interactions between the discriminator and the generator. We develop supporting theoretical analysis from the perspectives of differentiable games and stochastic approximation. Our experiments demonstrate that latent optimisation can significantly improve GAN training, obtaining state-of-the-art performance for the ImageNet (128 x 128) dataset. Our model achieves an Inception Score (IS) of 148 and an Frechet Inception Distance (FID) of 3.4, an improvement of 17% and 32% in IS and FID respectively, compared with the baseline BigGAN-deep model with the same architecture and number of parameters.
****
Hyper-SAGNN: a self-attention based graph neural network for hypergraphs
Ruochi Zhang,Yuesong Zou,Jian Ma
Graph representation learning for hypergraphs can be utilized to extract patterns among higher-order interactions that are critically important in many real world problems. Current approaches designed for hypergraphs, however, are unable to handle different types of hypergraphs and are typically not generic for various learning tasks. Indeed, models that can predict variable-sized heterogeneous hyperedges have not been available. Here we develop a new self-attention based graph neural network called Hyper-SAGNN applicable to homogeneous and heterogeneous hypergraphs with variable hyperedge sizes. We perform extensive evaluations on multiple datasets, including four benchmark network datasets and two single-cell Hi-C datasets in genomics. We demonstrate that Hyper-SAGNN significantly outperforms state-of-the-art methods on traditional tasks while also achieving great performance on a new task called outsider identification. We believe that Hyper-SAGNN will be useful for graph representation learning to uncover complex higher-order interactions in different applications.
****
A Neural Dirichlet Process Mixture Model for Task-Free Continual Learning
Soochan Lee,Junsoo Ha,Dongsu Zhang,Gunhee Kim
Despite the growing interest in continual learning, most of its contemporary works have been studied in a rather restricted setting where tasks are clearly distinguishable, and task boundaries are known during training. However, if our goal is to develop an algorithm that learns as humans do, this setting is far from realistic, and it is essential to develop a methodology that works in a task-free

manner. Meanwhile, among several branches of continual learning, expansion-based methods have the advantage of eliminating catastrophic forgetting by allocating new resources to learn new data. In this work, we propose an expansion-based approach for task-free continual learning. Our model, named Continual Neural Dirichlet Process Mixture (CN-DPM), consists of a set of neural network experts that are in charge of a subset of the data. CN-DPM expands the number of experts in a principled way under the Bayesian nonparametric framework. With extensive experiments, we show that our model successfully performs task-free continual learning for both discriminative and generative tasks such as image classification and image generation.

****

Global-Local Network for Learning Depth with Very Sparse Supervision

Antonio Loquercio,Alexey Dosovitskiy,Davide Scaramuzza

Natural intelligent agents learn to perceive the three dimensional structure of the world without training on large datasets and are unlikely to have the precise equations of projective geometry hard-wired in the brain. Such skill would also be valuable to artificial systems in order to avoid the expensive collection of labeled datasets, as well as tedious tuning required by methods based on multi-view geometry. Inspired by natural agents, who interact with the environment via visual and haptic feedback, this paper explores a new approach to learning depth from images and very sparse depth measurements, just a few pixels per image. To learn from such extremely sparse supervision, we introduce an appropriate inductive bias by designing a specialized global-local network architecture. Experiments on several datasets show that the proposed model can learn monocular dense depth estimation when trained with very sparse ground truth, even a single pixel per image. Moreover, we find that the global parameters extracted by the network are predictive of the metric agent motion.

****

CEB Improves Model Robustness

Ian Fischer,Alex A. Alemi

We demonstrate that the Conditional Entropy Bottleneck (CEB) can improve model robustness. CEB is an easy strategy to implement and works in tandem with data augmentation procedures. We report results of a large scale adversarial robustness study on CIFAR-10, as well as the IMAGENET-C Common Corruptions Benchmark.

****

Music Source Separation in the Waveform Domain

Alexandre Defossez,Nicolas Usunier,Leon Bottou,Francis Bach

Source separation for music is the task of isolating contributions, or stems, from different instruments recorded individually and arranged together to form a song.Such components include voice, bass, drums and any other accompaniments. While end-to-end models that directly generate the waveform are state-of-the-art in many audio synthesis problems, the best multi-instrument source separation models generate masks on the magnitude spectrum and achieve performances far above current end-to-end, waveform-to-waveform models. We present an in-depth analysis of a new architecture, which we will refer to as Demucs, based on a (transposed) convolutional autoencoder, with a bidirectional LSTM at the bottleneck layer and skip-connections as in U-Networks (Ronneberger et al., 2015). Compared to the state-of-the-art waveform-to-waveform model, Wave-U-Net (Stoller et al., 2018), the main features of our approach in addition of the bi-LSTM are the use of trans-posed convolution layers instead of upsampling-convolution blocks, the use of gated linear units, exponentially growing the number of channels with depth and a new careful initialization of the weights. Results on the MusDB dataset show that our architecture achieves a signal-to-distortion ratio (SDR) nearly 2.2 points higher than the best waveform-to-waveform competitor (from 3.2 to 5.4 SDR). This makes our model match the state-of-the-art performances on this dataset, bridging the performance gap between models that operate on the spectrogram and end-to-end approaches.

****

On Solving Minimax Optimization Locally: A Follow-the-Ridge Approach

Yuanhao Wang*,Guodong Zhang*,Jimmy Ba
Many tasks in modern machine learning can be formulated as finding equilibria in sequential games. In particular, two-player zero-sum sequential games, also known as minimax optimization, have received growing interest. It is tempting to apply gradient descent to solve minimax optimization given its popularity and success in supervised learning. However, it has been noted that naive application of gradient descent fails to find some local minimax and can converge to non-local-minimax points. In this paper, we propose Follow-the-Ridge (FR), a novel algorithm that provably converges to and only converges to local minimax. We show theoretically that the algorithm addresses the notorious rotational behaviour of gradient dynamics, and is compatible with preconditioning and positive momentum. Empirically, FR solves toy minimax problems and improves the convergence of GAN training compared to the recent minimax optimization algorithms.
****
Distributionally Robust Neural Networks
Shiori Sagawa*,Pang Wei Koh*,Tatsunori B. Hashimoto,Percy Liang
Overparameterized neural networks can be highly accurate on average on an i.i.d. test set, yet consistently fail on atypical groups of the data (e.g., by learning spurious correlations that hold on average but not in such groups). Distributionally robust optimization (DRO) allows us to learn models that instead minimize the worst-case training loss over a set of pre-defined groups. However, we find that naively applying group DRO to overparameterized neural networks fails: these models can perfectly fit the training data, and any model with vanishing average training loss also already has vanishing worst-case training loss. Instead, the poor worst-case performance arises from poor generalization on some groups. By coupling group DRO models with increased regularization---stronger-than-typical L2 regularization or early stopping---we achieve substantially higher worst-group accuracies, with 10-40 percentage point improvements on a natural language inference task and two image tasks, while maintaining high average accuracies. Our results suggest that regularization is important for worst-group generalization in the overparameterized regime, even if it is not needed for average generalization. Finally, we introduce a stochastic optimization algorithm for the group DRO setting and provide convergence guarantees for the new algorithm.

****
Kernel of CycleGAN as a principal homogeneous space
Nikita Moriakov,Jonas Adler,Jonas Teuwen
Unpaired image-to-image translation has attracted significant interest due to the invention of CycleGAN, a method which utilizes a combination of adversarial and cycle consistency losses to avoid the need for paired data. It is known that the CycleGAN problem might admit multiple solutions, and our goal in this paper is to analyze the space of exact solutions and to give perturbation bounds for approximate solutions. We show theoretically that the exact solution space is invariant with respect to automorphisms of the underlying probability spaces, and, furthermore, that the group of automorphisms acts freely and transitively on the space of exact solutions. We examine the case of zero pure CycleGAN loss first in its generality, and, subsequently, expand our analysis to approximate solutions for extended CycleGAN loss where identity loss term is included. In order to demonstrate that these results are applicable, we show that under mild conditions nontrivial smooth automorphisms exist. Furthermore, we provide empirical evidence that neural networks can learn these automorphisms with unexpected and unwanted results. We conclude that finding optimal solutions to the CycleGAN loss does not necessarily lead to the envisioned result in image-to-image translation tasks and that underlying hidden symmetries can render the result useless.
****
Don't Use Large Mini-batches, Use Local SGD
Tao Lin,Sebastian U. Stich,Kumar Kshitij Patel,Martin Jaggi
Mini-batch stochastic gradient methods (SGD) are state of the art for distributed training of deep neural networks.
Drastic increases in the mini-batch sizes have lead to key efficiency and scalab

ility gains in recent years.
However, progress faces a major roadblock, as models trained with large batches often do not generalize well, i.e. they do not show good accuracy on new data. As a remedy, we propose a \emph{post-local} SGD and show that it significantly improves the generalization performance compared to large-batch training on standard benchmarks while enjoying the same efficiency (time-to-accuracy) and scalability. We further provide an extensive study of the communication efficiency vs. performance trade-offs associated with a host of \emph{local SGD} variants.

****
Provable robustness against all adversarial l_p-perturbations for $p\geq 1$
Francesco Croce,Matthias Hein
In recent years several adversarial attacks and defenses have been proposed. Often seemingly robust models turn out to be non-robust when more sophisticated attacks are used. One way out of this dilemma are provable robustness guarantees. While provably robust models for specific l_p-perturbation models have been developed, we show that they do not come with any guarantee against other l_q-perturbations. We propose a new regularization scheme, MMR-Universal, for ReLU networks which enforces robustness wrt l_1- \textit{and} l_∞-perturbations and show how that leads to the first provably robust models wrt any l_p-norm for $p\geq 1$.
****
Model Based Reinforcement Learning for Atari
Łukasz Kaiser,Mohammad Babaeizadeh,Piotr Miłos,Błażej Osiński,Roy H Campbell,Konrad Czechowski,Dumitru Erhan,Chelsea Finn,Piotr Kozakowski,Sergey Levine,Afroz Mohiuddin,Ryan Sepassi,George Tucker,Henryk Michalewski
Model-free reinforcement learning (RL) can be used to learn effective policies for complex tasks, such as Atari games, even from image observations. However, this typically requires very large amounts of interaction -- substantially more, in fact, than a human would need to learn the same games. How can people learn so quickly? Part of the answer may be that people can learn how the game works and predict which actions will lead to desirable outcomes. In this paper, we explore how video prediction models can similarly enable agents to solve Atari games with fewer interactions than model-free methods. We describe Simulated Policy Learning (SimPLe), a complete model-based deep RL algorithm based on video prediction models and present a comparison of several model architectures, including a novel architecture that yields the best results in our setting. Our experiments evaluate SimPLe on a range of Atari games in low data regime of 100k interactions between the agent and the environment, which corresponds to two hours of real-time play. In most games SimPLe outperforms state-of-the-art model-free algorithms, in some games by over an order of magnitude.
****
Generating Multi-Sentence Abstractive Summaries of Interleaved Texts
Sanjeev Kumar Karn,Francine Chen,Yan-Ying Chen,Ulli Waltinger,Hinrich Schütze
In multi-participant postings, as in online chat conversations, several conversations or topic threads may take place concurrently. This leads to difficulties for readers reviewing the postings in not only following discussions but also in quickly identifying their essence. A two-step process, disentanglement of interleaved posts followed by summarization of each thread, addresses the issue, but disentanglement errors are propagated to the summarization step, degrading the overall performance. To address this, we propose an end-to-end trainable encoder-decoder network for summarizing interleaved posts. The interleaved posts are encoded hierarchically, i.e., word-to-word (words in a post) followed by post-to-post (posts in a channel). The decoder also generates summaries hierarchically, thread-to-thread (generate thread representations) followed by word-to-word (i.e., generate summary words). Additionally, we propose a hierarchical attention mechanism for interleaved text. Overall, our end-to-end trainable hierarchical framework enhances performance over a sequence to sequence framework by 8-10% on multiple synthetic interleaved texts datasets.
****

On Universal Equivariant Set Networks
Nimrod Segol,Yaron Lipman
Using deep neural networks that are either invariant or equivariant to permutations in order to learn functions on unordered sets has become prevalent. The most popular, basic models are DeepSets (Zaheer et al. 2017) and PointNet (Qi et al. 2017). While known to be universal for approximating invariant functions, DeepSets and PointNet are not known to be universal when approximating equivariant set functions. On the other hand, several recent equivariant set architectures have been proven equivariant universal (Sannai et al. 2019, Keriven and Peyre 2019), however these models either use layers that are not permutation equivariant (in the standard sense) and/or use higher order tensor variables which are less practical. There is, therefore, a gap in understanding the universality of popular equivariant set models versus theoretical ones.
■■■
In this paper we close this gap by proving that: (i) PointNet is not equivariant universal; and (ii) adding a single linear transmission layer makes PointNet universal. We call this architecture PointNetST and argue it is the simplest permutation equivariant universal model known to date. Another consequence is that DeepSets is universal, and also PointNetSeg, a popular point cloud segmentation network (used e.g., in Qi et al. 2017) is universal.
■■
The key theoretical tool used to prove the above results is an explicit characterization of all permutation equivariant polynomial layers. Lastly, we provide numerical experiments validating the theoretical results and comparing different permutation equivariant models.
****
OPTIMAL BINARY QUANTIZATION FOR DEEP NEURAL NETWORKS
Hadi Pouransari,Oncel Tuzel
Quantizing weights and activations of deep neural networks results in significant improvement in inference efficiency at the cost of lower accuracy. A source of the accuracy gap between full precision and quantized models is the quantization error.
In this work, we focus on the binary quantization, in which values are mapped to -1 and 1. We introduce several novel quantization algorithms: optimal 2-bits, optimal ternary, and greedy. Our quantization algorithms can be implemented efficiently on the hardware using bitwise operations. We present proofs to show that our proposed methods are optimal, and also provide empirical error analysis. We conduct experiments on the ImageNet dataset and show a reduced accuracy gap when using the proposed optimal quantization algorithms.
****
Deep End-to-end Unsupervised Anomaly Detection
Li Tangqing,Wang Zheng,Liu Siying,Daniel Lin Wen-Yan
This paper proposes a novel method to detect anomalies in large datasets under a fully unsupervised setting. The key idea behind our algorithm is to learn the representation underlying normal data. To this end, we leverage the latest clustering
technique suitable for handling high dimensional data. This hypothesis provides a reliable starting point for normal data selection. We train an autoencoder from the normal data subset, and iterate between hypothesizing normal candidate subset
based on clustering and representation learning. The reconstruction error from the learned autoencoder serves as a scoring function to assess the normality of the data. Experimental results on several public benchmark datasets show that the proposed method outperforms state-of-the-art unsupervised techniques and is comparable to semi-supervised techniques in most cases.
****
Tensor Decompositions for Temporal Knowledge Base Completion
Timothée Lacroix,Guillaume Obozinski,Nicolas Usunier
Most algorithms for representation learning and link prediction in relational data have been designed for static data. However, the data they are applied to usu

ally evolves with time, such as friend graphs in social networks or user interac
tions with items in recommender systems. This is also the case for knowledge bas
es, which contain facts such as (US, has president, B. Obama, [2009-2017]) that
are valid only at certain points in time. For the problem of link prediction und
er temporal constraints, i.e., answering queries of the form (US, has president,
 ?, 2012), we propose a solution inspired by the canonical decomposition of tens
ors of order 4.
We introduce new regularization schemes and present an extension of ComplEx that
 achieves state-of-the-art performance. Additionally, we propose a new dataset f
or knowledge base completion constructed from Wikidata, larger than previous ben
chmarks by an order of magnitude, as a new reference for evaluating temporal and
 non-temporal link prediction methods.
****
CloudLSTM: A Recurrent Neural Model for Spatiotemporal Point-cloud Stream Foreca
sting
Chaoyun Zhang,Marco Fiore,Iain Murray,Paul Patras
This paper introduces CloudLSTM, a new branch of recurrent neural models tailore
d to forecasting over data streams generated by geospatial point-cloud sources.
We design a Dynamic Point-cloud Convolution (D-Conv) operator as the core compon
ent of CloudLSTMs, which performs convolution directly over point-clouds and ext
racts local spatial features from sets of neighboring points that surround diffe
rent elements of the input. This operator maintains the permutation invariance o
f sequence-to-sequence learning frameworks, while representing neighboring corre
lations at each time step -- an important aspect in spatiotemporal predictive le
arning. The D-Conv operator resolves the grid-structural data requirements of ex
isting spatiotemporal forecasting models and can be easily plugged into traditio
nal LSTM architectures with sequence-to-sequence learning and attention mechanis
ms.
    We apply our proposed architecture to two representative, practical use case
s that involve point-cloud streams, i.e. mobile service traffic forecasting and
air quality indicator forecasting. Our results, obtained with real-world dataset
s collected in diverse scenarios for each use case, show that CloudLSTM delivers
 accurate long-term predictions, outperforming a variety of neural network model
s.
****
Neural Approximation of an Auto-Regressive Process through Confidence Guided Sam
pling
YoungJoon Yoo,Sanghyuk Chun,Jaejun Yoo,Sangdoo Yun,Jung Woo Ha
We propose a generic confidence-based approximation that can be plugged in and s
implify an auto-regressive generation process with a proved convergence. We firs
t assume that the priors of future samples can be generated in an independently
and identically distributed (i.i.d.) manner using an efficient predictor. Given
the past samples and future priors, the mother AR model can post-process the pri
ors while the accompanied confidence predictor decides whether the current sampl
e needs a resampling or not. Thanks to the i.i.d. assumption, the post-processin
g can update each sample in a parallel way, which remarkably accelerates the mot
her model. Our experiments on different data domains including sequences and ima
ges show that the proposed method can successfully capture the complex structure
s of the data and generate the meaningful future samples with lower computationa
l cost while preserving the sequential relationship of the data.}
****
Network Randomization: A Simple Technique for Generalization in Deep Reinforceme
nt Learning
Kimin Lee,Kibok Lee,Jinwoo Shin,Honglak Lee
Deep reinforcement learning (RL) agents often fail to generalize to unseen envir
onments (yet semantically similar to trained agents), particularly when they are
 trained on high-dimensional state spaces, such as images. In this paper, we pro
pose a simple technique to improve a generalization ability of deep RL agents by
 introducing a randomized (convolutional) neural network that randomly perturbs
input observations. It enables trained agents to adapt to new domains by learnin

g robust features invariant across varied and randomized environments. Furthermo
re, we consider an inference method based on the Monte Carlo approximation to re
duce the variance induced by this randomization. We demonstrate the superiority
of our method across 2D CoinRun, 3D DeepMind Lab exploration and 3D robotics con
trol tasks: it significantly outperforms various regularization and data augment
ation methods for the same purpose.
****

Stochastic Latent Residual Video Prediction
Jean-Yves Franceschi,Edouard Delasalles,Mickael Chen,Sylvain Lamprier,Patrick Ga
llinari
Video prediction is a challenging task: models have to account for the inherent
uncertainty of the future. Most works in the literature are based on stochastic
image-autoregressive recurrent networks, raising several performance and applica
bility issues. An alternative is to use fully latent temporal models which untie
 frame synthesis and dynamics. However, no such model for video prediction has b
een proposed in the literature yet, due to design and training difficulties. In
this paper, we overcome these difficulties by introducing a novel stochastic tem
poral model. It is based on residual updates of a latent state, motivated by dis
cretization schemes of differential equations. This first-order principle natura
lly models video dynamics as it allows our simpler, lightweight, interpretable,
latent model to outperform prior state-of-the-art methods on challenging dataset
s.
****

AlignNet: Self-supervised Alignment Module
Antonia Creswell,Luis Piloto,David Barrett,Kyriacos Nikiforou,David Raposo,Marta
 Garnelo,Peter Battaglia,Murray Shanahan
The natural world consists of objects that we perceive as persistent in space an
d time, even though these objects appear, disappear and reappear in our field of
 view as we move. This can be attributed to our notion of object persistence --
our knowledge that objects typically continue to exist, even if we can no longer
 see them -- and our ability to track objects. Drawing inspiration from the psyc
hology literature on `sticky indices', we propose the AlignNet, a model that lea
rns to assign unique indices to new objects when they first appear and reassign
the index to subsequent instances of that object. By introducing a persistent ob
ject-based memory, the AlignNet may be used to keep track of objects across time
, even if they disappear and reappear later. We implement the AlignNet as a grap
h network applied to a bipartite graph, in which the input nodes are objects fro
m two sets that we wish to align. The network is trained to predict the edges wh
ich connect two instances of the same object across sets. The model is also capa
ble of identifying when there are no matches and dealing with these cases. We pe
rform experiments to show the model's ability to deal with the appearance, disap
pearance and reappearance of objects. Additionally, we demonstrate how a persist
ent object-based memory can help solve question-answering problems in a partiall
y observable environment.
****

Learning with Protection: Rejection of Suspicious Samples under Adversarial Envi
ronment
Masahiro Kato,Yoshihiro Fukuhara,Hirokatsu Kataoka,Shigeo Morishima
We propose a novel framework for avoiding the misclassification of data by using
 a framework of learning with rejection and adversarial examples. Recent develop
ments in machine learning have opened new opportunities for industrial innovatio
ns such as self-driving cars. However, many machine learning models are vulnerab
le to adversarial attacks and industrial practitioners are concerned about accid
ents arising from misclassification. To avoid critical misclassifications, we de
fine a sample that is likely to be mislabeled as a suspicious sample. Our main i
dea is to apply a framework of learning with rejection and adversarial examples
to assist in the decision making for such suspicious samples. We propose two fra
meworks, learning with rejection under adversarial attacks and learning with pro
tection. Learning with rejection under adversarial attacks is a naive extension
of the learning with rejection framework for handling adversarial examples. Lear

ning with protection is a practical application of learning with rejection under adversarial attacks. This algorithm transforms the original multi-class classification problem into a binary classification for a specific class, and we reject suspicious samples to protect a specific label. We demonstrate the effectiveness of the proposed method in experiments.
****

QXplore: Q-Learning Exploration by Maximizing Temporal Difference Error
Riley Simmons-Edler,Ben Eisner,Daniel Yang,Anthony Bisulco,Eric Mitchell,Sebastian Seung,Daniel Lee
A major challenge in reinforcement learning is exploration, especially when reward landscapes are sparse. Several recent methods provide an intrinsic motivation to explore by directly encouraging agents to seek novel states. A potential disadvantage of pure state novelty-seeking behavior is that unknown states are treated equally regardless of their potential for future reward. In this paper, we propose an exploration objective using the temporal difference error experienced on extrinsic rewards as a secondary reward signal for exploration in deep reinforcement learning. Our objective yields novelty-seeking in the absence of extrinsic reward, while accelerating exploration of reward-relevant states in sparse (but nonzero) reward landscapes. This objective draws inspiration from dopaminergic pathways in the brain that influence animal behavior. We implement the objective with an adversarial Q-learning method in which Q and Qx are the action-value functions for extrinsic and secondary rewards, respectively. Secondary reward is given by the absolute value of the TD-error of Q. Training is off-policy, based on a replay buffer containing a mix of trajectories sampled using Q and Qx. We characterize performance on a set of continuous control benchmark tasks, and demonstrate comparable or faster convergence on all tasks when compared with other state-of-the-art exploration methods.
****

Walking the Tightrope: An Investigation of the Convolutional Autoencoder Bottleneck
Ilja Manakov,Markus Rohm,Volker Tresp
In this paper, we present an in-depth investigation of the convolutional autoencoder (CAE) bottleneck.
Autoencoders (AE), and especially their convolutional variants, play a vital role in the current deep learning toolbox.
Researchers and practitioners employ CAEs for a variety of tasks, ranging from outlier detection and compression to transfer and representation learning.
Despite their widespread adoption, we have limited insight into how the bottleneck shape impacts the emergent properties of the CAE.
We demonstrate that increased height and width of the bottleneck drastically improves generalization, which in turn leads to better performance of the latent codes in downstream transfer learning tasks.
The number of channels in the bottleneck, on the other hand, is secondary in importance.
Furthermore, we show empirically, that, contrary to popular belief, CAEs do not learn to copy their input, even when the bottleneck has the same number of neurons as there are pixels in the input.
Copying does not occur, despite training the CAE for 1,000 epochs on a tiny (~ 600 images) dataset.
We believe that the findings in this paper are directly applicable and will lead to improvements in models that rely on CAEs.
****

Partial Simulation for Imitation Learning
Nir Baram,Shie Mannor
Model-based imitation learning methods require full knowledge of the transition kernel for policy evaluation. In this work, we introduce the Expert Induced Markov Decision Process (eMDP) model as a formulation of solving imitation problems using Reinforcement Learning (RL), when only partial knowledge about the transition kernel is available. The idea of eMDP is to replace the unknown transition kernel with a synthetic kernel that: a) simulate the transition of state componen

ts for which the transition kernel is known (s_r), and b) extract from demonstra tions the state components for which the kernel is unknown (s_u). The next state is then stitched from the two components: s={s_r,s_u}. We describe in detail th e recipe for building an eMDP and analyze the errors caused by its synthetic ker nel. Our experiments include imitation tasks in multiplayer games, where the age nt has to imitate one expert in the presence of other experts for whom we cannot provide a transition model. We show that combining a policy gradient algorithm with our model achieves superior performance compared to the simulation-free alt ernative.
****

Few-shot Learning by Focusing on Differences
Muhammad Rizki Maulana,Lee Wee Sun
Few-shot classification may involve differentiating data that belongs to a diffe rent level of labels granularity. Compounded by the fact that the number of avai lable labeled examples are scarce in the novel classification set, relying solel y on the loss function to implicitly guide the classifier to separate data based on its label might not be enough; few-shot classifier needs to be very biased t o perform well. In this paper, we propose a model that incorporates a simple pri or: focusing on differences by building a dissimilar set of class representation s. The model treats a class representation as a vector and removes its component that is shared among closely related class representatives. It does so through the combination of learned attention and vector orthogonalization. Our model wor ks well on our newly introduced dataset, Hierarchical-CIFAR, that contains diffe rent level of labels granularity. It also substantially improved the performance on fine-grained classification dataset, CUB; whereas staying competitive on sta ndard benchmarks such as mini-Imagenet, Omniglot, and few-shot dataset derived f rom CIFAR.
****

Robustness Verification for Transformers
Zhouxing Shi,Huan Zhang,Kai-Wei Chang,Minlie Huang,Cho-Jui Hsieh
Robustness verification that aims to formally certify the prediction behavior of neural networks has become an important tool for understanding model behavior a nd obtaining safety guarantees. However, previous methods can usually only handl e neural networks with relatively simple architectures. In this paper, we consid er the robustness verification problem for Transformers. Transformers have compl ex self-attention layers that pose many challenges for verification, including c ross-nonlinearity and cross-position dependency, which have not been discussed i n previous works. We resolve these challenges and develop the first robustness v erification algorithm for Transformers. The certified robustness bounds computed by our method are significantly tighter than those by naive Interval Bound Prop agation. These bounds also shed light on interpreting Transformers as they consi stently reflect the importance of different words in sentiment analysis.
****

Fantastic Generalization Measures and Where to Find Them
Yiding Jiang*,Behnam Neyshabur*,Hossein Mobahi,Dilip Krishnan,Samy Bengio
Generalization of deep networks has been intensely researched in recent years, r esulting in a number of theoretical bounds and empirically motivated measures. H owever, most papers proposing such measures only study a small set of models, le aving open the question of whether these measures are truly useful in practice. We present the first large scale study of generalization bounds and measures in deep networks. We train over two thousand CIFAR-10 networks with systematic chan ges in important hyper-parameters. We attempt to uncover potential causal relati onships between each measure and generalization, by using rank correlation coeff icient and its modified forms. We analyze the results and show that some of the studied measures are very promising for further research.
****

Nesterov Accelerated Gradient and Scale Invariance for Adversarial Attacks
Jiadong Lin,Chuanbiao Song,Kun He,Liwei Wang,John E. Hopcroft
Deep learning models are vulnerable to adversarial examples crafted by applying human-imperceptible perturbations on benign inputs. However, under the black-box

setting, most existing adversaries often have a poor transferability to attack other defense models. In this work, from the perspective of regarding the adversarial example generation as an optimization process, we propose two new methods to improve the transferability of adversarial examples, namely Nesterov Iterative Fast Gradient Sign Method (NI-FGSM) and Scale-Invariant attack Method (SIM). NI-FGSM aims to adapt Nesterov accelerated gradient into the iterative attacks so as to effectively look ahead and improve the transferability of adversarial examples. While SIM is based on our discovery on the scale-invariant property of deep learning models, for which we leverage to optimize the adversarial perturbations over the scale copies of the input images so as to avoid "overfitting" on the white-box model being attacked and generate more transferable adversarial examples. NI-FGSM and SIM can be naturally integrated to build a robust gradient-based attack to generate more transferable adversarial examples against the defense models. Empirical results on ImageNet dataset demonstrate that our attack methods exhibit higher transferability and achieve higher attack success rates than state-of-the-art gradient-based attacks.

****

Learning De-biased Representations with Biased Representations

Hyojin Bahng,Sanghyuk Chun,Sangdoo Yun,Jaegul Choo,Seong Joon Oh

Many machine learning algorithms are trained and evaluated by splitting data from a single source into training and test sets. While such focus on in-distribution learning scenarios has led interesting advances, it has not been able to tell if models are relying on dataset biases as shortcuts for successful prediction (e.g., using snow cues for recognising snowmobiles). Such biased models fail to generalise when the bias shifts to a different class. The cross-bias generalisation problem has been addressed by de-biasing training data through augmentation or re-sampling, which are often prohibitive due to the data collection cost (e.g., collecting images of snowmobile on a desert) and the difficulty of quantifying or expressing biases in the first place. In this work, we propose a novel framework to train a de-biased representation by encouraging it to be different from a set of representations that are biased by design. This tactic is feasible in many scenarios where it is much easier to define a set of biased representations than to define and quantify bias. Our experiments and analyses show that our method discourages models from taking bias shortcuts, resulting in improved performances on de-biased test data.

****

Weakly Supervised Disentanglement with Guarantees

Rui Shu,Yining Chen,Abhishek Kumar,Stefano Ermon,Ben Poole

Learning disentangled representations that correspond to factors of variation in real-world data is critical to interpretable and human-controllable machine learning. Recently, concerns about the viability of learning disentangled representations in a purely unsupervised manner has spurred a shift toward the incorporation of weak supervision. However, there is currently no formalism that identifies when and how weak supervision will guarantee disentanglement. To address this issue, we provide a theoretical framework to assist in analyzing the disentanglement guarantees (or lack thereof) conferred by weak supervision when coupled with learning algorithms based on distribution matching. We empirically verify the guarantees and limitations of several weak supervision methods (restricted labeling, match-pairing, and rank-pairing), demonstrating the predictive power and usefulness of our theoretical framework.

****

Imagining the Latent Space of a Variational Auto-Encoders

Zezhen Zeng,Jonathon Hare,Adam Prügel-Bennett

Variational Auto-Encoders (VAEs) are designed to capture compressible information about a dataset. As a consequence the information stored in the latent space is seldom sufficient to reconstruct a particular image. To help understand the type of information stored in the latent space we train a GAN-style decoder constrained to produce images that the VAE encoder will map to the same region of latent space. This allows us to ''imagine'' the information captured in the latent space. We argue that this is necessary to make a VAE into a truly generative

model. We use our GAN to visualise the latent space of a standard VAE and of a β-VAE.
****

A Copula approach for hyperparameter transfer learning
David Salinas,Huibin Shen,Valerio Perrone
Bayesian optimization (BO) is a popular methodology to tune the hyperparameters of expensive black-box functions. Despite its success, standard BO focuses on a single task at a time and is not designed to leverage information from related functions, such as tuning performance metrics of the same algorithm across multiple datasets. In this work, we introduce a novel approach to achieve transfer learning across different datasets as well as different metrics. The main idea is to regress the mapping from hyperparameter to metric quantiles with a semi-parametric Gaussian Copula distribution, which provides robustness against different scales or outliers that can occur in different tasks. We introduce two methods to leverage this estimation: a Thompson sampling strategy as well as a Gaussian Copula process using such quantile estimate as a prior. We show that these strategies can combine the estimation of multiple metrics such as runtime and accuracy, steering the optimization toward cheaper hyperparameters for the same level of accuracy. Experiments on an extensive set of hyperparameter tuning tasks demonstrate significant improvements over state-of-the-art methods.
****

THE EFFECT OF ADVERSARIAL TRAINING: A THEORETICAL CHARACTERIZATION
Mingyang Yi,Huishuai Zhang,Wei Chen,Zhi-Ming Ma,Tie-Yan Liu
It has widely shown that adversarial training (Madry et al., 2018) is effective in defending adversarial attack empirically. However, the theoretical understanding of the difference between the solution of adversarial training and that of standard training is limited. In this paper, we characterize the solution of adversarial training for linear classification problem for a full range of adversarial radius ". Specifically, we show that if the data themselves are ”-strongly linearly-separable”, adversarial
training with radius smaller than " converges to the hard margin solution of SVM with a faster rate than standard training. If the data themselves are not ”-strongly linearly-separable”, we show that adversarial training with radius " is stable to outliers while standard training is not. Moreover, we prove that the classifier returned by adversarial training with a large radius " has low confidence in each data point. Experiments corroborate our theoretical finding well.
****

Provenance detection through learning transformation-resilient watermarking
Jamie Hayes,Krishnamurthy Dvijotham,Yutian Chen,Sander Dieleman,Pushmeet Kohli,Norman Casagrande
Advancements in deep generative models have made it possible to synthesize images, videos and audio signals that are hard to distinguish from natural signals, creating opportunities for potential abuse of these capabilities. This motivates the problem of tracking the provenance of signals, i.e., being able to determine the original source of a signal. Watermarking the signal at the time of signal creation is a potential solution, but current techniques are brittle and watermark detection mechanisms can easily be bypassed by doing some post-processing (cropping images, shifting pitch in the audio etc.). In this paper, we introduce ReSWAT (Resilient Signal Watermarking via Adversarial Training), a framework for learning transformation-resilient watermark detectors that are able to detect a watermark even after a signal has been through several post-processing transformations. Our detection method can be applied to domains with continuous data representations such as images, videos or sound signals. Experiments on watermarking image and audio signals show that our method can reliably detect the provenance of a synthetic signal, even if the signal has been through several post-processing transformations, and improve upon related work in this setting. Furthermore, we show that for specific kinds of transformations (perturbations bounded in the ℓ_2 norm), we can even get formal guarantees on the ability of our model to detect the watermark. We provide qualitative examples of watermarked image and audio samples in the anonymous code submission link.

****
Regulatory Focus: Promotion and Prevention Inclinations in Policy Search
Lanxin Lei,Zhizhong Li,Xiaoyang Li,Cong Qiu,Dahua Lin
The estimation of advantage is crucial for a number of reinforcement learning al gorithms, as it directly influences the choices of future paths. In this work, w e propose a family of estimates based on the order statistics over the path ense mble, which allows one to flexibly drive the learning process in a promotion foc us or prevention focus. On top of this formulation, we systematically study the impacts of different regulatory focuses. Our findings reveal that regulatory foc us, when chosen appropriately, can result in significant benefits. In particular , for the environments with sparse rewards, promotion focus would lead to more e fficient exploration of the policy space; while for those where individual actio ns can have critical impacts, prevention focus is preferable. On various benchma rks, including MuJoCo continuous control, Terrain locomotion, Atari games, and s parse-reward environments, the proposed schemes consistently demonstrate improve ment over mainstream methods, not only accelerating the learning process but als o obtaining substantial performance gains.
****
Fairness with Wasserstein Adversarial Networks
serrurier Mathieu,Loubes Jean-Michel,Edouard Pauwels
Quantifying, enforcing and implementing fairness emerged as a major topic in mac hine learning. We investigate these questions in the context of deep learning. O ur main algorithmic and theoretical tool is the computational estimation of simi larities between probability, ``\`a la Wasserstein'', using adversarial networks . This idea is flexible enough to investigate different fairness constrained lea rning tasks, which we model by specifying properties of the underlying data gene rative process. The first setting considers bias in the generative model which s hould be filtered out. The second model is related to the presence of nuisance v ariables in the observations producing an unwanted bias for the learning task. For both models, we devise a learning algorithm based on approximation of Wasser stein distances using adversarial networks. We provide formal arguments describi ng the fairness enforcing properties of these algorithm in relation with the und erlying fairness generative processes. Finally we perform experiments, both on s ynthetic and real world data, to demonstrate empirically the superiority of our approach compared to state of the art fairness algorithms as well as concurrent GAN type adversarial architectures based on Jensen divergence.
****
Diagonal Graph Convolutional Networks with Adaptive Neighborhood Aggregation
Jie Zhang,Yuxiao Dong,Jie Tang
Graph convolutional networks (GCNs) and their variants have generalized deep lea rning methods into non-Euclidean graph data, bringing a substantial improvement on many graph mining tasks. In this paper, we revisit the mathematical foundatio n of GCNs and study how to extend their representation capacity. We discover tha t their performance can be improved with an adaptive neighborhood aggregation st ep. The core idea is to adaptively scale the output signal for each node and aut omatically train a suitable nonlinear encoder for the input signal. In this work , we present a new method named Diagonal Graph Convolutional Networks (DiagGCN) based on this idea. Importantly, one of the adaptive aggregation techniques—the permutations of diagonal matrices—used in DiagGCN offers a flexible framework to design GCNs and in fact, some of the most expressive GCNs, e.g., the graph atte ntion network, can be reformulated as a particular instance of our model. Standa rd experiments on open graph benchmarks show that our proposed framework can con sistently improve the graph classification accuracy when compared to state-of-th e-art baselines.
****
Discrepancy Ratio: Evaluating Model Performance When Even Experts Disagree on th e Truth
Igor Lovchinsky,Alon Daks,Israel Malkin,Pouya Samangouei,Ardavan Saeedi,Yang Liu ,Swami Sankaranarayanan,Tomer Gafner,Ben Sternlieb,Patrick Maher,Nathan Silberma n

In most machine learning tasks unambiguous ground truth labels can easily be acq
uired. However, this luxury is often not afforded to many high-stakes, real-worl
d scenarios such as medical image interpretation, where even expert human annota
tors typically exhibit very high levels of disagreement with one another. While
prior works have focused on overcoming noisy labels during training, the questio
n of how to evaluate models when annotators disagree about ground truth has rema
ined largely unexplored. To address this, we propose the discrepancy ratio: a no
vel, task-independent and principled framework for validating machine learning m
odels in the presence of high label noise. Conceptually, our approach evaluates
a model by comparing its predictions to those of human annotators, taking into a
ccount the degree to which annotators disagree with one another. While our appro
ach is entirely general, we show that in the special case of binary classificati
on, our proposed metric can be evaluated in terms of simple, closed-form express
ions that depend only on aggregate statistics of the labels and not on any indiv
idual label. Finally, we demonstrate how this framework can be used effectively
to validate machine learning models using two real-world tasks from medical imag
ing. The discrepancy ratio metric reveals what conventional metrics do not: that
 our models not only vastly exceed the average human performance, but even excee
d the performance of the best human experts in our datasets.
****

The Dual Information Bottleneck

Zoe Piran,Naftali Tishby

The Information-Bottleneck (IB) framework suggests a general characterization of
 optimal representations in learning, and deep learning in particular. It is bas
ed on the optimal trade off between the representation complexity and accuracy,
both of which are quantified by mutual information. The problem is solved by alt
ernating projections between the encoder and decoder of the representation, whic
h can be performed locally at each representation level. The framework, however,
 has practical drawbacks, in that mutual information is notoriously difficult to
 handle at high dimension, and only has closed form solutions in special cases.
Further, because it aims to extract representations which are minimal sufficient
 statistics of the data with respect to the desired label, it does not necessari
ly optimize the actual prediction of unseen labels. Here we present a formal du
al problem to the IB which has several interesting properties. By switching the
order in the KL-divergence between the representation decoder and data, the opti
mal decoder becomes the geometric rather than the arithmetic mean of the input p
oints. While providing a good approximation to the original IB, it also preserve
s the form of exponential families, and optimizes the mutual information on the
predicted label rather than the desired one. We also analyze the critical points
 of the dualIB and discuss their importance for the quality of this approach.
****

Deep Auto-Deferring Policy for Combinatorial Optimization

Sungsoo Ahn,Younggyo Seo,Jinwoo Shin

Designing efficient algorithms for combinatorial optimization appears ubiquitous
ly in various scientific fields. Recently, deep reinforcement learning (DRL) fra
meworks have gained considerable attention as a new approach: they can automatic
ally learn the design of a good solver without using any sophisticated knowledge
 or hand-crafted heuristic specialized for the target problem. However, the numb
er of stages (until reaching the final solution) required by existing DRL solver
s is proportional to the size of the input graph, which hurts their scalability
to large-scale instances. In this paper, we seek to resolve this issue by propos
ing a novel design of DRL's policy, coined auto-deferring policy (ADP), automati
cally stretching or shrinking its decision process. Specifically, it decides whe
ther to finalize the value of each vertex at the current stage or defer to deter
mine it at later stages. We apply the proposed ADP framework to the maximum inde
pendent set (MIS) problem, a prototype of NP-complete problems, under various sc
enarios. Our experimental results demonstrate significant improvement of ADP ove
r the current state-of-the-art DRL scheme in terms of computational efficiency a
nd approximation quality. The reported performance of our generic DRL scheme is
also comparable with that of the state-of-the-art solvers specialized for MIS, e

.g., ADP outperforms them for some graphs with millions of vertices.
****
Towards trustworthy predictions from deep neural networks with fast adversarial calibration
Christian Tomani,Florian Buettner
To facilitate a wide-spread acceptance of AI systems guiding decision making in real-world applications, trustworthiness of deployed models is key. That is, it is crucial for predictive models to be uncertainty-aware and yield well-calibrated (and thus trustworthy) predictions for both in-domain samples as well as under domain shift. Recent efforts to account for predictive uncertainty include post-processing steps for trained neural networks, Bayesian neural networks as well as alternative non-Bayesian approaches such as ensemble approaches and evidential deep learning. Here, we propose an efficient yet general modelling approach for obtaining well-calibrated, trustworthy probabilities for samples obtained after a domain shift. We introduce a new training strategy combining an entropy-encouraging loss term with an adversarial calibration loss term and demonstrate that this results in well-calibrated and technically trustworthy predictions for a wide range of perturbations. We comprehensively evaluate previously proposed approaches on different data modalities, a large range of data sets, network architectures and perturbation strategies and observe that our modelling approach substantially outperforms existing state-of-the-art approaches, yielding well-calibrated predictions for both in-domain and out-of domain samples.
****
Abductive Commonsense Reasoning
Chandra Bhagavatula,Ronan Le Bras,Chaitanya Malaviya,Keisuke Sakaguchi,Ari Holtzman,Hannah Rashkin,Doug Downey,Wen-tau Yih,Yejin Choi
Abductive reasoning is inference to the most plausible explanation. For example, if Jenny finds her house in a mess when she returns from work, and remembers that she left a window open, she can hypothesize that a thief broke into her house and caused the mess, as the most plausible explanation. While abduction has long been considered to be at the core of how people interpret and read between the lines in natural language (Hobbs et al., 1988), there has been relatively little research in support of abductive natural language inference and generation. We present the first study that investigates the viability of language-based abductive reasoning. We introduce a challenge dataset, ART, that consists of over 20k commonsense narrative contexts and 200k explanations. Based on this dataset, we conceptualize two new tasks – (i) Abductive NLI: a multiple-choice question answering task for choosing the more likely explanation, and (ii) Abductive NLG: a conditional generation task for explaining given observations in natural language. On Abductive NLI, the best model achieves 68.9% accuracy, well below human performance of 91.4%. On Abductive NLG, the current best language generators struggle even more, as they lack reasoning capabilities that are trivial for humans. Our analysis leads to new insights into the types of reasoning that deep pre-trained language models fail to perform—despite their strong performance on the related but more narrowly defined task of entailment NLI—pointing to interesting avenues for future research.
****
Variance Reduction With Sparse Gradients
Melih Elibol,Lihua Lei,Michael I. Jordan
Variance reduction methods such as SVRG and SpiderBoost use a mixture of large and small batch gradients to reduce the variance of stochastic gradients. Compared to SGD, these methods require at least double the number of operations per update to model parameters. To reduce the computational cost of these methods, we introduce a new sparsity operator: The random-top-k operator. Our operator reduces computational complexity by estimating gradient sparsity exhibited in a variety of applications by combining the top-k operator and the randomized coordinate descent operator. With this operator, large batch gradients offer an extra benefit beyond variance reduction: A reliable estimate of gradient sparsity. Theoretically, our algorithm is at least as good as the best algorithm (SpiderBoost), and further excels in performance whenever the random-top-k operator captures grad

ient sparsity. Empirically, our algorithm consistently outperforms SpiderBoost u
sing various models on various tasks including image classification, natural lan
guage processing, and sparse matrix factorization. We also provide empirical evi
dence to support the intuition behind our algorithm via a simple gradient entrop
y computation, which serves to quantify gradient sparsity at every iteration.
****

BlockSwap: Fisher-guided Block Substitution for Network Compression on a Budget
Jack Turner,Elliot J. Crowley,Michael O'Boyle,Amos Storkey,Gavin Gray
The desire to map neural networks to varying-capacity devices has led to the dev
elopment of a wealth of compression techniques, many of which involve replacing
standard convolutional blocks in a large network with cheap alternative blocks.
However, not all blocks are created equally; for a required compute budget there
 may exist a potent combination of many different cheap blocks, though exhaustiv
ely searching for such a combination is prohibitively expensive. In this work, w
e develop BlockSwap: a fast algorithm for choosing networks with interleaved blo
ck types by passing a single minibatch of training data through randomly initial
ised networks and gauging their Fisher potential. These networks can then be use
d as students and distilled with the original large network as a teacher. We dem
onstrate the effectiveness of the chosen networks across CIFAR-10 and ImageNet f
or classification, and COCO for detection, and provide a comprehensive ablation
study of our approach. BlockSwap quickly explores possible block configurations
using a simple architecture ranking system, yielding highly competitive networks
 in orders of magnitude less time than most architecture search techniques (e.g.
 under 5 minutes on a single GPU for CIFAR-10).
****

RNA Secondary Structure Prediction By Learning Unrolled Algorithms
Xinshi Chen,Yu Li,Ramzan Umarov,Xin Gao,Le Song
In this paper, we propose an end-to-end deep learning model, called E2Efold, for
 RNA secondary structure prediction which can effectively take into account the
inherent constraints in the problem. The key idea of E2Efold is to directly pred
ict the RNA base-pairing matrix, and use an unrolled algorithm for constrained p
rogramming as the template for deep architectures to enforce constraints. With c
omprehensive experiments on benchmark datasets, we demonstrate the superior perf
ormance of E2Efold: it predicts significantly better structures compared to prev
ious SOTA (especially for pseudoknotted structures), while being as efficient as
 the fastest algorithms in terms of inference time.
****

Learning transport cost from subset correspondence
Ruishan Liu,Akshay Balsubramani,James Zou
Learning to align multiple datasets is an important problem with many applicatio
ns, and it is especially useful when we need to integrate multiple experiments o
r correct for confounding. Optimal transport (OT) is a principled approach to al
ign datasets, but a key challenge in applying OT is that we need to specify a c
ost function that accurately captures how the two datasets are related. Reliable
 cost functions are typically not available and practitioners often resort to us
ing hand-crafted or Euclidean cost even if it may not be appropriate. In this wo
rk, we investigate how to learn the cost function using a small amount of side i
nformation which is often available. The side information we consider captures s
ubset correspondence---i.e. certain subsets of points in the two data sets are k
nown to be related. For example, we may have some images labeled as cars in both
 datasets; or we may have a common annotated cell type in single-cell data from
two batches. We develop an end-to-end optimizer (OT-SI) that differentiates thro
ugh the Sinkhorn algorithm and effectively learns the suitable cost function fro
m side information. On systematic experiments in images, marriage-matching and s
ingle-cell RNA-seq, our method substantially outperform state-of-the-art benchma
rks.
****

Semi-Supervised Few-Shot Learning with a Controlled Degree of Task-Adaptive Cond
itioning
Sung Whan Yoon,Jun Seo,Jaekyun Moon

Few-shot learning aims to handle previously unseen tasks using only a small amou nt of new training data. In preparing (or meta-training) a few-shot learner, how ever, massive labeled data are necessary. In the real world, unfortunately, labe led data are expensive and/or scarce. In this work, we propose a few-shot learne r that can work well under the semi-supervised setting where a large portion of training data is unlabeled. Our method employs explicit task-conditioning in whi ch unlabeled sample clustering for the current task takes place in a new project ion space different from the embedding feature space. The conditioned clustering space is linearly constructed so as to quickly close the gap between the class centroids for the current task and the independent per-class reference vectors m eta-trained across tasks. In a more general setting, our method introduces a con cept of controlling the degree of task-conditioning for meta-learning: the amoun t of task-conditioning varies with the number of repetitive updates for the clus tering space. During each update, the soft labels of the unlabeled samples estim ated in the conditioned clustering space are used to update the class averages i n the original embedded space, which in turn are used to reconstruct the cluster ing space. Extensive simulation results based on the miniImageNet and tieredImag eNet datasets show state-of-the-art semi-supervised few-shot classification perf ormance of the proposed method. Simulation results also indicate that the propos ed task-adaptive clustering shows graceful degradation with a growing number of distractor samples, i.e., unlabeled samples coming from outside the candidate cl asses.

****

Detecting Noisy Training Data with Loss Curves

Geoff Pleiss,Tianyi Zhang,Ethan R. Elenberg,Kilian Q. Weinberger

This paper introduces a new method to discover mislabeled training samples and t o mitigate their impact on the training process of deep networks. At the heart o f our algorithm lies the Area Under the Loss (AUL) statistic, which can be easil y computed for each sample in the training set. We show that the AUL can use tra ining dynamics to differentiate between (clean) samples that benefit from genera lization and (mislabeled) samples that need to be “memorized”. We demonstrate th at the estimated AUL score conditioned on clean vs. noisy is approximately Gauss ian distributed and can be well estimated with a simple Gaussian Mixture Model (GMM). The resulting GMM provides us with mixing coefficients that reveal the per centage of mislabeled samples in a data set as well as probability estimates tha t each individual training sample is mislabeled. We show that these probability estimates can be used to down-weight suspicious training samples and successfull y alleviate the damaging impact of label noise. We demonstrate on the CIFAR10/10 0 datasets that our proposed approach is significantly more accurate and consist ent across model architectures than all prior work.

****

Near-Zero-Cost Differentially Private Deep Learning with Teacher Ensembles

Lichao Sun,Yingbo Zhou,Jia Li,Richard Socher,Philip S. Yu,Caiming Xiong

Ensuring the privacy of sensitive data used to train modern machine learning mod els is of paramount importance in many areas of practice. One approach to study these concerns is through the lens of differential privacy. In this framework, p rivacy guarantees are generally obtained by perturbing models in such a way that specifics of data used to train the model are made ambiguous. A particular inst ance of this approach is through a ``teacher-student'' model, wherein the teache r, who owns the sensitive data, provides the student with useful, but noisy, inf ormation, hopefully allowing the student model to perform well on a given task w ithout access to particular features of the sensitive data. Because stronger pri vacy guarantees generally involve more significant noising on the part of the te acher, deploying existing frameworks fundamentally involves a trade-off between utility and privacy guarantee. One of the most important techniques used in prev ious work involves an ensemble of teacher models, which return information to a student based on a noisy voting procedure. In this work, we propose a novel vot ing mechanism, which we call an Immutable Noisy ArgMax, that, under certain cond itions, can bear very large random noising from the teacher without affecting th e useful information transferred to the student. Our mechanisms improve over the

state-of-the-art methods on all measures, and scale to larger tasks with both higher utility and stronger privacy ($\epsilon \approx 0$).
****
Neural Network Out-of-Distribution Detection for Regression Tasks
Geoff Pleiss,Amauri Souza,Joseph Kim,Boyi Li,Kilian Q. Weinberger
Neural network out-of-distribution (OOD) detection aims to identify when a model is unable to generalize to new inputs, either due to covariate shift or anomalous data. Most existing OOD methods only apply to classification tasks, as they assume a discrete set of possible predictions. In this paper, we propose a method for neural network OOD detection that can be applied to regression problems. We demonstrate that the hidden features for in-distribution data can be described by a highly concentrated, low dimensional distribution. Therefore, we can model these in-distribution features with an extremely simple generative model, such as a Gaussian mixture model (GMM) with 4 or fewer components. We demonstrate on several real-world benchmark data sets that GMM-based feature detection achieves state-of-the-art OOD detection results on several regression tasks. Moreover, this approach is simple to implement and computationally efficient.
****
Rényi Fair Inference
Sina Baharlouei,Maher Nouiehed,Ahmad Beirami,Meisam Razaviyayn
Machine learning algorithms have been increasingly deployed in critical automated decision-making systems that directly affect human lives. When these algorithms are solely trained to minimize the training/test error, they could suffer from systematic discrimination against individuals based on their sensitive attributes, such as gender or race. Recently, there has been a surge in machine learning society to develop algorithms for fair machine learning.
In particular, several adversarial learning procedures have been proposed to impose fairness. Unfortunately, these algorithms either can only impose fairness up to linear dependence between the variables, or they lack computational convergence guarantees. In this paper, we use Rényi correlation as a measure of fairness of machine learning models and develop a general training framework to impose fairness. In particular, we propose a min-max formulation which balances the accuracy and fairness when solved to optimality. For the case of discrete sensitive attributes, we suggest an iterative algorithm with theoretical convergence guarantee for solving the proposed min-max problem. Our algorithm and analysis are then specialized to fair classification and fair clustering problems. To demonstrate the performance of the proposed Rényi fair inference framework in practice, we compare it with well-known existing methods on several benchmark datasets. Experiments indicate that the proposed method has favorable empirical performance against state-of-the-art approaches.
****
Reject Illegal Inputs: Scaling Generative Classifiers with Supervised Deep Infomax
Xin WANG,SiuMing Yiu
Deep Infomax~(DIM) is an unsupervised representation learning framework by maximizing the mutual information between the inputs and the outputs of an encoder, while probabilistic constraints are imposed on the outputs. In this paper, we propose Supervised Deep InfoMax~(SDIM), which introduces supervised probabilistic constraints to the encoder outputs. The supervised probabilistic constraints are equivalent to a generative classifier on high-level data representations, where class conditional log-likelihoods of samples can be evaluated. Unlike other works building generative classifiers with conditional generative models, SDIMs scale on complex datasets, and can achieve comparable performance with discriminative counterparts. With SDIM, we could perform \emph{classification with rejection}.
Instead of always reporting a class label, SDIM only makes predictions when test samples' largest logits surpass some pre-chosen thresholds, otherwise they will be deemed as out of the data distributions, and be rejected. Our experiments show that SDIM with rejection policy can effectively reject illegal inputs including out-of-distribution samples and adversarial examples.

****
Lean Images for Geo-Localization
Moti Kadosh,Yael Moses,Ariel Shamir
Most computer vision tasks use textured images. In this paper we consider the geo-localization task - finding the pose of a camera in a large 3D scene from a single lean image, i.e. an image with no texture. We aim to experimentally explore whether texture and correlation between nearby images are necessary in a CNN-based solution for this task. Our results may give insight to the role of geometry (as opposed to textures) in a CNN-based geo-localization solution. Lean images are projections of a simple 3D model of a city. They contain solely information that relates to the geometry of the scene viewed (edges, faces, or relative depth). We find that the network is capable of estimating the camera pose from lean images for a relatively large number of locations (order of hundreds of thousands of images). The main contributions of this paper are: (i) demonstrating the power of CNNs for recovering camera pose using lean images; and (ii) providing insight into the role of geometry in the CNN learning process;
****
WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia
Holger Schwenk,Vishrav Chaudhary,Shuo Sun,Hongyu Gong,Francisco Guzmán
We present an approach based on multilingual sentence embeddings to automatically extract parallel sentences from the content of Wikipedia articles in 85 languages, including several dialects or low-resource languages. We do not limit the extraction process to alignments with English, but systematically consider all possible language pairs. In total, we are able to extract 135M parallel sentences for 1620 different language pairs, out of which only 34M are aligned with English. This corpus of parallel sentences is freely available (URL anonymized)

To get an indication on the quality of the extracted bitexts, we train neural MT baseline systems on the mined data only for 1886 languages pairs, and evaluate them on the TED corpus, achieving strong BLEU scores for many language pairs. The WikiMatrix bitexts seem to be particularly interesting to train MT systems between distant languages without the need to pivot through English.
****
Deep Lifetime Clustering
S Chandra Mouli,Leonardo Teixeira,Jennifer Neville,Bruno Ribeiro
The goal of lifetime clustering is to develop an inductive model that maps subjects into K clusters according to their underlying (unobserved) lifetime distribution. We introduce a neural-network based lifetime clustering model that can find cluster assignments by directly maximizing the divergence between the empirical lifetime distributions of the clusters. Accordingly, we define a novel clustering loss function over the lifetime distributions (of entire clusters) based on a tight upper bound of the two-sample Kuiper test p-value. The resultant model is robust to the modeling issues associated with the unobservability of termination signals, and does not assume proportional hazards. Our results in real and synthetic datasets show significantly better lifetime clusters (as evaluated by C-index, Brier Score, Logrank score and adjusted Rand index) as compared to competing approaches.
****
Towards Understanding the Transferability of Deep Representations
Hong Liu,Mingsheng Long,Jianmin Wang,Michael I. Jordan
Deep neural networks trained on a wide range of datasets demonstrate impressive transferability. Deep features appear general in that they are applicable to many datasets and tasks. Such property is in prevalent use in real-world applications. A neural network pretrained on large datasets, such as ImageNet, can significantly boost generalization and accelerate training if fine-tuned to a smaller target dataset. Despite its pervasiveness, few effort has been devoted to uncovering the reason of transferability in deep feature representations. This paper tries to understand transferability from the perspectives of improved generalization, optimization and the feasibility of transferability. We demonstrate that 1) Transferred models tend to find flatter minima, since their weight matrices stay

close to the original flat region of pretrained parameters when transferred to a similar target dataset; 2) Transferred representations make the loss landscape more favorable with improved Lipschitzness, which accelerates and stabilizes training substantially. The improvement largely attributes to the fact that the principal component of gradient is suppressed in the pretrained parameters, thus stabilizing the magnitude of gradient in back-propagation. 3) The feasibility of transferability is related to the similarity of both input and label. And a surprising discovery is that the feasibility is also impacted by the training stages in that the transferability first increases during training, and then declines. We further provide a theoretical analysis to verify our observations.
****

Meta Dropout: Learning to Perturb Latent Features for Generalization
Hae Beom Lee,Taewook Nam,Eunho Yang,Sung Ju Hwang
A machine learning model that generalizes well should obtain low errors on unseen test examples. Thus, if we know how to optimally perturb training examples to account for test examples, we may achieve better generalization performance. However, obtaining such perturbation is not possible in standard machine learning frameworks as the distribution of the test data is unknown. To tackle this challenge, we propose a novel regularization method, meta-dropout, which learns to perturb the latent features of training examples for generalization in a meta-learning framework. Specifically, we meta-learn a noise generator which outputs a multiplicative noise distribution for latent features, to obtain low errors on the test instances in an input-dependent manner. Then, the learned noise generator can perturb the training examples of unseen tasks at the meta-test time for improved generalization. We validate our method on few-shot classification datasets, whose results show that it significantly improves the generalization performance of the base model, and largely outperforms existing regularization methods such as information bottleneck, manifold mixup, and information dropout.
****

Adversarial AutoAugment
Xinyu Zhang,Qiang Wang,Jian Zhang,Zhao Zhong
Data augmentation (DA) has been widely utilized to improve generalization in training deep neural networks. Recently, human-designed data augmentation has been gradually replaced by automatically learned augmentation policy. Through finding the best policy in well-designed search space of data augmentation, AutoAugment (Cubuk et al., 2019) can significantly improve validation accuracy on image classification tasks. However, this approach is not computationally practical for large-scale problems. In this paper, we develop an adversarial method to arrive at a computationally-affordable solution called Adversarial AutoAugment, which can simultaneously optimize target related object and augmentation policy search loss. The augmentation policy network attempts to increase the training loss of a target network through generating adversarial augmentation policies, while the target network can learn more robust features from harder examples to improve the generalization. In contrast to prior work, we reuse the computation in target network training for policy evaluation, and dispense with the retraining of the target network. Compared to AutoAugment, this leads to about 12x reduction in computing cost and 11x shortening in time overhead on ImageNet. We show experimental results of our approach on CIFAR-10/CIFAR-100, ImageNet, and demonstrate significant performance improvements over state-of-the-art. On CIFAR-10, we achieve a top-1 test error of 1.36%, which is the currently best performing single model. On ImageNet, we achieve a leading performance of top-1 accuracy 79.40% on ResNet-50 and 80.00% on ResNet-50-D without extra data.
****

When Robustness Doesn’t Promote Robustness: Synthetic vs. Natural Distribution Shifts on ImageNet
Rohan Taori,Achal Dave,Vaishaal Shankar,Nicholas Carlini,Benjamin Recht,Ludwig Schmidt
We conduct a large experimental comparison of various robustness metrics for image classification. The main question of our study is to what extent current synthetic robustness interventions (lp-adversarial examples, noise corruptions, etc.

) promote robustness under natural distribution shifts occurring in real data. T
o this end, we evaluate 147 ImageNet models under 199 different evaluation setti
ngs. We find that no current robustness intervention improves robustness on natu
ral distribution shifts beyond a baseline given by standard models without a rob
ustness intervention. The only exception is the use of larger training datasets,
 which provides a small increase in robustness on one natural distribution shift
. Our results indicate that robustness improvements on real data may require new
 methodology and more evaluations on natural distribution shifts.
****

Understanding Why Neural Networks Generalize Well Through GSNR of Parameters
Jinlong Liu,Yunzhi Bai,Guoqing Jiang,Ting Chen,Huayan Wang
As deep neural networks (DNNs) achieve tremendous success across many applicatio
n domains, researchers tried to explore in many aspects on why they generalize w
ell. In this paper, we provide a novel perspective on these issues using the gra
dient signal to noise ratio (GSNR) of parameters during training process of DNNs
. The GSNR of a parameter is simply defined as the ratio between its gradient's
squared mean and variance, over the data distribution. Based on several approxim
ations, we establish a quantitative relationship between model parameters' GSNR
and the generalization gap. This relationship indicates that larger GSNR during
training process leads to better generalization performance. Futher, we show tha
t, different from that of shallow models (e.g. logistic regression, support vect
or machines), the gradient descent optimization dynamics of DNNs naturally produ
ces large GSNR during training, which is probably the key to DNNs’ remarkable ge
neralization ability.
****

State-only Imitation with Transition Dynamics Mismatch
Tanmay Gangwani,Jian Peng
Imitation Learning (IL) is a popular paradigm for training agents to achieve com
plicated goals by leveraging expert behavior, rather than dealing with the hards
hips of designing a correct reward function. With the environment modeled as a M
arkov Decision Process (MDP), most of the existing IL algorithms are contingent
on the availability of expert demonstrations in the same MDP as the one in which
 a new imitator policy is to be learned. This is uncharacteristic of many real-l
ife scenarios where discrepancies between the expert and the imitator MDPs are c
ommon, especially in the transition dynamics function. Furthermore, obtaining ex
pert actions may be costly or infeasible, making the recent trend towards state-
only IL (where expert demonstrations constitute only states or observations) eve
r so promising. Building on recent adversarial imitation approaches that are mot
ivated by the idea of divergence minimization, we present a new state-only IL al
gorithm in this paper. It divides the overall optimization objective into two su
bproblems by introducing an indirection step and solves the subproblems iterativ
ely. We show that our algorithm is particularly effective when there is a transi
tion dynamics mismatch between the expert and imitator MDPs, while the baseline
IL methods suffer from performance degradation. To analyze this, we construct se
veral interesting MDPs by modifying the configuration parameters for the MuJoCo
locomotion tasks from OpenAI Gym.
****

Measuring and Improving the Use of Graph Information in Graph Neural Networks
Yifan Hou,Jian Zhang,James Cheng,Kaili Ma,Richard T. B. Ma,Hongzhi Chen,Ming-Cha
ng Yang
Graph neural networks (GNNs) have been widely used for representation learning o
n graph data. However, there is limited understanding on how much performance GN
Ns actually gain from graph data. This paper introduces a context-surrounding GN
N framework and proposes two smoothness metrics to measure the quantity and qual
ity of information obtained from graph data. A new, improved GNN model, called C
S-GNN, is then devised to improve the use of graph information based on the smoo
thness values of a graph. CS-GNN is shown to achieve better performance than exi
sting methods in different types of real graphs.
****

Meta-Learning by Hallucinating Useful Examples

Yu-Xiong Wang,Yuki Uchiyama,Martial Hebert,Karteek Alahari
Learning to hallucinate additional examples has recently been shown as a promising direction to address few-shot learning tasks, which aim to learn novel concepts from very few examples. The hallucination process, however, is still far from generating effective samples for learning. In this work, we investigate two important requirements for the hallucinator --- (i) precision: the generated examples should lead to good classifier performance, and (ii) collaboration: both the hallucinator and the classification component need to be trained jointly. By integrating these requirements as novel loss functions into a general meta-learning with hallucination framework, our model-agnostic PrecisE Collaborative hAlluciNator (PECAN) facilitates data hallucination to improve the performance of new classification tasks. Extensive experiments demonstrate state-of-the-art performance on competitive miniImageNet and ImageNet based few-shot benchmarks in various scenarios.
****

Pixel Co-Occurence Based Loss Metrics for Super Resolution Texture Recovery
Ying Da Wang,Pawel Swietojanski,Ryan T Armstrong,Peyman Mostaghimi
Single Image Super Resolution (SISR) has significantly improved with Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs), often achieving order of magnitude better pixelwise accuracies (distortions) and state-of-the-art perceptual accuracy. Due to the stochastic nature of GAN reconstruction and the ill-posed nature of the problem, perceptual accuracy tends to correlate inversely with pixelwise accuracy which is especially detrimental to SISR, where preservation of original content is an objective. GAN stochastics can be guided by intermediate loss functions such as the VGG featurewise loss, but these features are typically derived from biased pre-trained networks. Similarly, measurements of perceptual quality such as the human Mean Opinion Score (MOS) and no-reference measures have issues with pre-trained bias. The spatial relationships between pixel values can be measured without bias using the Grey Level Co-occurence Matrix (GLCM), which was found to match the cardinality and comparative value of the MOS while reducing subjectivity and automating the analytical process. In this work, the GLCM is also directly used as a loss function to guide the generation of perceptually accurate images based on spatial collocation of pixel values. We compare GLCM based loss against scenarios where (1) no intermediate guiding loss function, and (2) the VGG feature function are used. Experimental validation is carried on X-ray images of rock samples, characterised by significant number of high frequency texture features. We find GLCM-based loss to result in images with higher pixelwise accuracy and better perceptual scores.
****

A Latent Morphology Model for Open-Vocabulary Neural Machine Translation
Duygu Ataman,Wilker Aziz,Alexandra Birch
Translation into morphologically-rich languages challenges neural machine translation (NMT) models with extremely sparse vocabularies where atomic treatment of surface forms is unrealistic. This problem is typically addressed by either pre-processing words into subword units or performing translation directly at the level of characters. The former is based on word segmentation algorithms optimized using corpus-level statistics with no regard to the translation task. The latter learns directly from translation data but requires rather deep architectures. In this paper, we propose to translate words by modeling word formation through a hierarchical latent variable model which mimics the process of morphological inflection. Our model generates words one character at a time by composing two latent representations: a continuous one, aimed at capturing the lexical semantics, and a set of (approximately) discrete features, aimed at capturing the morphosyntactic function, which are shared among different surface forms. Our model achieves better accuracy in translation into three morphologically-rich languages than conventional open-vocabulary NMT methods, while also demonstrating a better generalization capacity under low to mid-resource settings.
****

Sample-Based Point Cloud Decoder Networks
Erich Merrill,Alan Fern

Point clouds are a flexible and ubiquitous way to represent 3D objects with arbi
trary resolution and precision. Previous work has shown that adapting encoder ne
tworks to match the semantics of their input point clouds can significantly impr
ove their effectiveness over naive feedforward alternatives. However, the vast m
ajority of work on point-cloud decoders are still based on fully-connected netwo
rks that map shape representations to a fixed number of output points. In this w
ork, we investigate decoder architectures that more closely match the semantics
of variable sized point clouds. Specifically, we study sample-based point-cloud
decoders that map a shape representation to a point feature distribution, allowi
ng an arbitrary number of sampled features to be transformed into individual out
put points. We develop three sample-based decoder architectures and compare thei
r performance to each other and show their improved effectiveness over feedforwa
rd architectures. In addition, we investigate the learned distributions to gain
insight into the output transformation. Our work is available as an extensible s
oftware platform to reproduce these results and serve as a baseline for future w
ork.
****

AUGMENTED POLICY GRADIENT METHODS FOR EFFICIENT REINFORCEMENT LEARNING
Kai Lagemann,Gregor Roering,Christoph Henke,Rene Vossen,Frank Hees
We propose a new mixture of model-based and model-free reinforcement learning
(RL) algorithms that combines the strengths of both RL methods. Our goal is to r
educe the sample complexity of model-free approaches utilizing fictitious trajec
tory
rollouts performed on a learned dynamics model to improve the data efficiency of
policy gradient methods while maintaining the same asymptotic behaviour. We
suggest to use a special type of uncertainty quantification by a stochastic dyna
mics
model in which the next state prediction is randomly drawn from the distribution
predicted by the dynamics model. As a result, the negative effect of exploiting
erroneously optimistic regions in the dynamics model is addressed by next state
predictions based on an uncertainty aware ensemble of dynamics models. The
influence of the ensemble of dynamics models on the policy update is controlled
by adjusting the number of virtually performed rollouts in the next iteration ac
cording to the ratio of the real and virtual total reward. Our approach, which w
e
call Model-Based Policy Gradient Enrichment (MBPGE), is tested on a collection o
f benchmark tests including simulated robotic locomotion. We compare our
approach to plain model-free algorithms and a model-based one. Our evaluation
shows that MBPGE leads to higher learning rates in an early training stage and a
n
improved asymptotic behaviour.
****

BETANAS: Balanced Training and selective drop for Neural Architecture Search
Muyuan Fang,Qiang Wang,Jian Zhang,Zhao Zhong
Automatic neural architecture search techniques are becoming increasingly import
ant in machine learning area recently. Especially, weight sharing methods have
shown remarkable potentials on searching good network architectures with few com
putational resources. However, existing weight sharing methods mainly suffer lim
itations on searching strategies: these methods either uniformly train all netwo
rk paths to convergence which introduces conflicts between branches and wastes a
 large amount of computation on unpromising candidates, or selectively train bra
nches with different frequency which leads to unfair evaluation and comparison a
mong paths. To address these issues, we propose a novel neural architecture sear
ch method with balanced training strategy to ensure fair comparisons and a sele
ctive drop mechanism to reduces conflicts among candidate paths. The experimenta
l results show that our proposed method can achieve a leading performance of 79.
0% on ImageNet under mobile settings, which outperforms other state-of-the-art m
ethods in both accuracy and efficiency.
****

Connecting the Dots Between MLE and RL for Sequence Prediction

Bowen Tan,Zhiting Hu,Zichao Yang,Ruslan Salakhutdinov,Eric Xing
Sequence prediction models can be learned from example sequences with a variety of training algorithms. Maximum likelihood learning is simple and efficient, yet can suffer from compounding error at test time.
Reinforcement learning such as policy gradient addresses the issue but can have prohibitively poor exploration efficiency. A rich set of other algorithms, such as data noising, RAML, and softmax policy gradient, have also been developed from different perspectives.
In this paper, we present a formalism of entropy regularized policy optimization, and show that the apparently distinct algorithms, including MLE, can be reformulated as special instances of the formulation. The difference between them is characterized by the reward function and two weight hyperparameters.
The unifying interpretation enables us to systematically compare the algorithms side-by-side, and gain new insights into the trade-offs of the algorithm design. The new perspective also leads to an improved approach that dynamically interpolates among the family of algorithms, and learns the model in a scheduled way. Experiments on machine translation, text summarization, and game imitation learning demonstrate superiority of the proposed approach.
****
Universal Approximation with Certified Networks
Maximilian Baader,Matthew Mirman,Martin Vechev
Training neural networks to be certifiably robust is critical to ensure their safety against adversarial attacks. However, it is currently very difficult to train a neural network that is both accurate and certifiably robust. In this work we take a step towards addressing this challenge. We prove that for every continuous function f, there exists a network n such that:
(i) n approximates f arbitrarily close, and (ii) simple interval bound propagation of a region B through n yields a result that is arbitrarily close to the optimal output of f on B. Our result can be seen as a Universal Approximation Theorem for interval-certified ReLU networks. To the best of our knowledge, this is the first work to prove the existence of accurate, interval-certified networks.
****
Explain Your Move: Understanding Agent Actions Using Specific and Relevant Feature Attribution
Nikaash Puri,Sukriti Verma,Piyush Gupta,Dhruv Kayastha,Shripad Deshmukh,Balaji Krishnamurthy,Sameer Singh
As deep reinforcement learning (RL) is applied to more tasks, there is a need to visualize and understand the behavior of learned agents. Saliency maps explain agent behavior by highlighting the features of the input state that are most relevant for the agent in taking an action. Existing perturbation-based approaches to compute saliency often highlight regions of the input that are not relevant to the action taken by the agent. Our proposed approach, SARFA (Specific and Relevant Feature Attribution), generates more focused saliency maps by balancing two aspects (specificity and relevance) that capture different desiderata of saliency. The first captures the impact of perturbation on the relative expected reward of the action to be explained. The second downweighs irrelevant features that alter the relative expected rewards of actions other than the action to be explained. We compare SARFA with existing approaches on agents trained to play board games (Chess and Go) and Atari games (Breakout, Pong and Space Invaders). We show through illustrative examples (Chess, Atari, Go), human studies (Chess), and automated evaluation methods (Chess) that SARFA generates saliency maps that are more interpretable for humans than existing approaches. For the code release and demo videos, see: https://nikaashpuri.github.io/sarfa-saliency/.
****
Learning to Balance: Bayesian Meta-Learning for Imbalanced and Out-of-distribution Tasks
Hae Beom Lee,Hayeon Lee,Donghyun Na,Saehoon Kim,Minseop Park,Eunho Yang,Sung Ju Hwang
While tasks could come with varying the number of instances and classes in reali

stic settings, the existing meta-learning approaches for few-shot classification assume that number of instances per task and class is fixed. Due to such restriction, they learn to equally utilize the meta-knowledge across all the tasks, even when the number of instances per task and class largely varies. Moreover, they do not consider distributional difference in unseen tasks, on which the meta-knowledge may have less usefulness depending on the task relatedness. To overcome these limitations, we propose a novel meta-learning model that adaptively balances the effect of the meta-learning and task-specific learning within each task. Through the learning of the balancing variables, we can decide whether to obtain a solution by relying on the meta-knowledge or task-specific learning. We formulate this objective into a Bayesian inference framework and tackle it using variational inference. We validate our Bayesian Task-Adaptive Meta-Learning (Bayesian TAML) on two realistic task- and class-imbalanced datasets, on which it significantly outperforms existing meta-learning approaches. Further ablation study confirms the effectiveness of each balancing component and the Bayesian learning framework.
****

DyNet: Dynamic Convolution for Accelerating Convolution Neural Networks
Kane Zhang,Jian Zhang,Qiang Wang,Zhao Zhong
Convolution operator is the core of convolutional neural networks (CNNs) and occupies the most computation cost. To make CNNs more efficient, many methods have been proposed to either design lightweight networks or compress models. Although some efficient network structures have been proposed, such as MobileNet or ShuffleNet, we find that there still exists redundant information between convolution kernels. To address this issue, we propose a novel dynamic convolution method named \textbf{DyNet} in this paper, which can adaptively generate convolution kernels based on image contents. To demonstrate the effectiveness, we apply DyNet on multiple state-of-the-art CNNs. The experiment results show that DyNet can reduce the computation cost remarkably, while maintaining the performance nearly unchanged. Specifically, for ShuffleNetV2 (1.0), MobileNetV2 (1.0), ResNet18 and ResNet50, DyNet reduces 40.0%, 56.7%, 68.2% and 72.4% FLOPs respectively while the Top-1 accuracy on ImageNet only changes by +1.0%, -0.27%, -0.6% and -0.08%. Meanwhile, DyNet further accelerates the inference speed of MobileNetV2 (1.0), ResNet18 and ResNet50 by 1.87x,1.32x and 1.48x on CPU platform respectively. To verify the scalability, we also apply DyNet on segmentation task, the results show that DyNet can reduces 69.3% FLOPs while maintaining the Mean IoU on segmentation task.
****

Deep Symbolic Superoptimization Without Human Knowledge
Hui Shi,Yang Zhang,Xinyun Chen,Yuandong Tian,Jishen Zhao
Deep symbolic superoptimization refers to the task of applying deep learning methods to simplify symbolic expressions. Existing approaches either perform supervised training on human-constructed datasets that defines equivalent expression pairs, or apply reinforcement learning with human-defined equivalent trans-formation actions. In short, almost all existing methods rely on human knowledge to define equivalence, which suffers from large labeling cost and learning bias, because it is almost impossible to define and comprehensive equivalent set. We thus propose HISS, a reinforcement learning framework for symbolic super-optimization that keeps human outside the loop. HISS introduces a tree-LSTM encoder-decoder network with attention to ensure tractable learning. Our experiments show that HISS can discover more simplification rules than existing human-dependent methods, and can learn meaningful embeddings for symbolic expressions, which are indicative of equivalence.
****

Unsupervised domain adaptation with imputation
Matthieu Kirchmeyer,Patrick Gallinari,Alain Rakotomamonjy,Amin Mantrach
Motivated by practical applications, we consider unsupervised domain adaptation for classification problems, in the presence of missing data in the target domain. More precisely, we focus on the case where there is a domain shift between source and target domains, while some components of the target data are systematic

ally absent. We propose a way to impute non-stochastic missing data for a classi fication task by leveraging supervision from a complete source domain through do main adaptation. We introduce a single model performing joint domain adaptation, imputation and classification which is shown to perform well under various repr esentative divergence families (H-divergence, Optimal Transport). We perform exp eriments on two families of datasets: a classical digit classification benchmark commonly used in domain adaptation papers and real world digital advertising da tasets, on which we evaluate our model’s classification performance in an unsupe rvised setting. We analyze its behavior showing the benefit of explicitly imputi ng non-stochastic missing data jointly with domain adaptation.
****

Sample Efficient Policy Gradient Methods with Recursive Variance Reduction
Pan Xu,Felicia Gao,Quanquan Gu
Improving the sample efficiency in reinforcement learning has been a long-standi ng research problem. In this work, we aim to reduce the sample complexity of exi sting policy gradient methods. We propose a novel policy gradient algorithm call ed SRVR-PG, which only requires $O(1/\epsilon^{3/2})$\footnote{$O(\cdot)$ notati on hides constant factors.} episodes to find an ϵ-approximate stationar y point of the nonconcave performance function $J(\boldsymbol{\theta})$ (i.e., $ \boldsymbol{\theta}$ such that $\|\nabla J(\boldsymbol{\theta})\|_2^2\leq\epsilo n$). This sample complexity improves the existing result $O(1/\epsilon^{5/3})$ f or stochastic variance reduced policy gradient algorithms by a factor of $O(1/\e psilon^{1/6})$. In addition, we also propose a variant of SRVR-PG with parameter exploration, which explores the initial policy parameter from a prior probabili ty distribution. We conduct numerical experiments on classic control problems in reinforcement learning to validate the performance of our proposed algorithms.
****

A Generative Model for Molecular Distance Geometry
Gregor N. C. Simm,José Miguel Hernández-Lobato
Computing equilibrium states for many-body systems, such as molecules, is a long -standing challenge. In the absence of methods for generating statistically inde pendent samples, great computational effort is invested in simulating these syst ems using, for example, Markov chain Monte Carlo. We present a probabilistic mod el that generates such samples for molecules from their graph representations. O ur model learns a low-dimensional manifold that preserves the geometry of local atomic neighborhoods through a principled learning representation that is based on Euclidean distance geometry. We create a new dataset for molecular conformati on generation with which we show experimentally that our generative model achiev es state-of-the-art accuracy. Finally, we show how to use our model as a proposa l distribution in an importance sampling scheme to compute molecular properties.
****

Fast Task Inference with Variational Intrinsic Successor Features
Steven Hansen,Will Dabney,Andre Barreto,David Warde-Farley,Tom Van de Wiele,Volo dymyr Mnih
It has been established that diverse behaviors spanning the controllable subspac e of a Markov decision process can be trained by rewarding a policy for being di stinguishable from other policies. However, one limitation of this formulation i s the difficulty to generalize beyond the finite set of behaviors being explicit ly learned, as may be needed in subsequent tasks. Successor features provide an appealing solution to this generalization problem, but require defining the rewa rd function as linear in some grounded feature space. In this paper, we show tha t these two techniques can be combined, and that each method solves the other's primary limitation. To do so we introduce Variational Intrinsic Successor FeatuR es (VISR), a novel algorithm which learns controllable features that can be leve raged to provide enhanced generalization and fast task inference through the suc cessor features framework. We empirically validate VISR on the full Atari suite, in a novel setup wherein the rewards are only exposed briefly after a long unsu pervised phase. Achieving human-level performance on 12 games and beating all ba selines, we believe VISR represents a step towards agents that rapidly learn fro m limited feedback.

****
Certified Defenses for Adversarial Patches
Ping-yeh Chiang*,Renkun Ni*,Ahmed Abdelkader,Chen Zhu,Christoph Studor,Tom Goldstein
Adversarial patch attacks are among one of the most practical threat models against real-world computer vision systems. This paper studies certified and empirical defenses against patch attacks. We begin with a set of experiments showing that most existing defenses, which work by pre-processing input images to mitigate adversarial patches, are easily broken by simple white-box adversaries. Motivated by this finding, we propose the first certified defense against patch attacks, and propose faster methods for its training. Furthermore, we experiment with different patch shapes for testing, obtaining surprisingly good robustness transfer across shapes, and present preliminary results on certified defense against sparse attacks. Our complete implementation can be found on: https://github.com/Ping-C/certifiedpatchdefense.
****
Contrastive Representation Distillation
Yonglong Tian,Dilip Krishnan,Phillip Isola
  Often we wish to transfer representational knowledge from one neural network to another. Examples include distilling a large network into a smaller one, transferring knowledge from one sensory modality to a second, or ensembling a collection of models into a single estimator. Knowledge distillation, the standard approach to these problems, minimizes the KL divergence between the probabilistic outputs of a teacher and student network. We demonstrate that this objective ignores important structural knowledge of the teacher network. This motivates an alternative objective by which we train a student to capture significantly more information in the teacher's representation of the data. We formulate this objective as contrastive learning. Experiments demonstrate that our resulting new objective outperforms knowledge distillation on a variety of knowledge transfer tasks, including single model compression, ensemble distillation, and cross-modal transfer. When combined with knowledge distillation, our method sets a state of the art in many transfer tasks, sometimes even outperforming the teacher network.
****
Generating valid Euclidean distance matrices
Moritz Hoffmann,Frank Noe
Generating point clouds, e.g., molecular structures, in arbitrary rotations, translations, and enumerations remains a challenging task. Meanwhile, neural networks
utilizing symmetry invariant layers have been shown to be able to optimize their training objective in a data-efficient way. In this spirit, we present an architecture
which allows to produce valid Euclidean distance matrices, which by construction are already invariant under rotation and translation of the described object.
Motivated by the goal to generate molecular structures in Cartesian space, we use
this architecture to construct a Wasserstein GAN utilizing a permutation invariant critic network. This makes it possible to generate molecular structures in a one-shot fashion by producing Euclidean distance matrices which have a three-dimensional embedding.
****
Perturbations are not Enough: Generating Adversarial Examples with Spatial Distortions
He Zhao,Trung Le,Paul Montague,Olivier De Vel,Tamas Abraham,Dinh Phung
Deep neural network image classifiers are reported to be susceptible to adversarial evasion attacks, which use carefully crafted images created to mislead a classifier. Recently, various kinds of adversarial attack methods have been proposed, most of which focus on adding small perturbations to input images. Despite the success of existing approaches, the way to generate realistic adversarial images with small perturbations remains a challenging problem. In this paper, we aim to address this problem by proposing a novel adversarial method, which generate

s adversarial examples by imposing not only perturbations but also spatial disto rtions on input images, including scaling, rotation, shear, and translation. As humans are less susceptible to small spatial distortions, the proposed approach can produce visually more realistic attacks with smaller perturbations, able to deceive classifiers without affecting human predictions. We learn our method by amortized techniques with neural networks and generate adversarial examples effi ciently by a forward pass of the networks. Extensive experiments on attacking di fferent types of non-robustified classifiers and robust classifiers with defence show that our method has state-of-the-art performance in comparison with advanc ed attack parallels.

****

Information Theoretic Model Predictive Q-Learning

Mohak Bhardwaj,Ankur Handa,Dieter Fox,Byron Boots

Model-free Reinforcement Learning (RL) algorithms work well in sequential decisi on-making problems when experience can be collected cheaply and model-based RL i s effective when system dynamics can be modeled accurately. However, both of the se assumptions can be violated in real world problems such as robotics, where qu erying the system can be prohibitively expensive and real-world dynamics can be difficult to model accurately. Although sim-to-real approaches such as domain ra ndomization attempt to mitigate the effects of biased simulation, they can still suffer from optimization challenges such as local minima and hand-designed dist ributions for randomization, making it difficult to learn an accurate global val ue function or policy that directly transfers to the real world. In contrast to RL, Model Predictive Control (MPC) algorithms use a simulator to optimize a simp le policy class online, constructing a closed-loop controller that can effective ly contend with real-world dynamics. MPC performance is usually limited by facto rs such as model bias and the limited horizon of optimization. In this work, we present a novel theoretical connection between information theoretic MPC and ent ropy regularized RL and develop a Q-learning algorithm that can leverage biased models. We validate the proposed algorithm on sim-to-sim control tasks to demons trate the improvements over optimal control and reinforcement learning from scra tch. Our approach paves the way for deploying reinforcement learning algorithms on real-robots in a systematic manner.

****

On Predictive Information Sub-optimality of RNNs

Zhe Dong,Deniz Oktay,Ben Poole,Alexander A. Alemi

Certain biological neurons demonstrate a remarkable capability to optimally comp ress the history of sensory inputs while being maximally informative about the f uture. In this work, we investigate if the same can be said of artificial neuron s in recurrent neural networks (RNNs) trained with maximum likelihood. In experi ments on two datasets, restorative Brownian motion and a hand-drawn sketch datas et, we find that RNNs are sub-optimal in the information plane. Instead of optim ally compressing past information, they extract additional information that is n ot relevant for predicting the future. Overcoming this limitation may require al ternative training procedures and architectures, or objectives beyond maximum li kelihood estimation.

****

Model Inversion Networks for Model-Based Optimization

Aviral Kumar,Sergey Levine

In this work, we aim to solve data-driven optimization problems, where the goal is to find an input that maximizes an unknown score function given access to a d ataset of input, score pairs. Inputs may lie on extremely thin manifolds in high -dimensional spaces, making the optimization prone to falling-off the manifold. Further, evaluating the unknown function may be expensive, so the algorithm shou ld be able to exploit static, offline data. We propose model inversion networks (MINs) as an approach to solve such problems. Unlike prior work, MINs scale to e xtremely high-dimensional input spaces and can efficiently leverage offline logg ed datasets for optimization in both contextual and non-contextual settings. We show that MINs can also be extended to the active setting, commonly studied in p rior work, via a simple, novel and effective scheme for active data collection.

Our experiments show that MINs act as powerful optimizers on a range of contextu
al/non-contextual, static/active problems including optimization over images and
 protein designs and learning from logged bandit feedback.
****

Learning to Recognize the Unseen Visual Predicates
Defa Zhu,Si Liu,Wentao Jiang,Guanbin Li,Tianyi Wu,Guodong Guo
Visual relationship recognition models are limited in the ability to generalize
 from finite seen predicates to unseen ones. We propose a new problem setting nam
ed predicate zero-shot learning (PZSL): learning to recognize the predicates wit
hout training data. It is unlike the previous zero-shot learning problem on visu
al relationship recognition which learns to recognize the unseen relationship tr
iplets (<subject, predicate, object>) but requires all components (subject, pred
icate, and object) to be seen in the training set. For the PZSL problem, however
, the models are expected to recognize the diverse even unseen predicates, which
 is meaningful for many downstream high-level tasks, like visual question answer
ing, to handle complex scenes and open questions. The PZSL is a very challenging
 task since the predicates are very abstract and follow an extreme long-tail dis
tribution. To address the PZSL problem, we present a model that performs compati
bility learning leveraging the linguistic priors from the corpus and knowledge b
ase. An unbalanced sampled-softmax is further developed to tackle the extreme lo
ng-tail distribution of predicates. Finally, the experiments are conducted to an
alyze the problem and verify the effectiveness of our methods. The dataset and s
ource code will be released for further study.
****

Continuous Control with Contexts, Provably
Simon Du,Mengdi Wang,Ruosong Wang,Lin F. Yang
A fundamental challenge in artificially intelligence is to build an agent that g
eneralizes and adapts to unseen environments. A common strategy is to build a de
coder that takes a context of the unseen new environment and generates a policy.
 The current paper studies how to build a decoder for the fundamental continuous
 control environment, linear quadratic regulator (LQR), which can model a wide r
ange of real world physical environments. We present a simple algorithm for this
 problem, which uses upper confidence bound (UCB) to refine the estimate of the
decoder and balance the exploration-exploitation trade-off. Theoretically, our a
lgorithm enjoys a $\widetilde{O}\left(\sqrt{T}\right)$ regret bound in the onlin
e setting where T is the number of environments the agent played. This also im
plies after playing $\widetilde{O}\left(1/\epsilon^2\right)$ environments, the a
gent is able to transfer the learned knowledge to obtain an ϵ-suboptima
l policy for an unseen environment. To our knowledge, this is first provably eff
icient algorithm to build a decoder in the continuous control setting. While our
 main focus is theoretical, we also present experiments that demonstrate the eff
ectiveness of our algorithm.
****

Stabilizing Transformers for Reinforcement Learning
Emilio Parisotto,Francis Song,Jack Rae,Razvan Pascanu,Caglar Gulcehre,Siddhant J
ayakumar,Max Jaderberg,Raphaël Lopez Kaufman,Aidan Clark,Seb Noury,Matt Botvinic
k,Nicolas Heess,Raia Hadsell
Owing to their ability to both effectively integrate information over long time
 horizons and scale to massive amounts of data, self-attention architectures have
 recently shown breakthrough success in natural language processing (NLP), achie
ving state-of-the-art results in domains such as language modeling and machine t
ranslation. Harnessing the transformer's ability to process long time horizons o
f information could provide a similar performance boost in partially-observable
 reinforcement learning (RL) domains, but the large-scale transformers used in NL
P have yet to be successfully applied to the RL setting. In this work we demonst
rate that the standard transformer architecture is difficult to optimize, which
 was previously observed in the supervised learning setting but becomes especiall
y pronounced with RL objectives. We propose architectural modifications that sub
stantially improve the stability and learning speed of the original Transformer
 and XL variant. The proposed architecture, the Gated Transformer-XL (GTrXL), sur

passes LSTMs on challenging memory environments and achieves state-of-the-art results on the multi-task DMLab-30 benchmark suite, exceeding the performance of an external memory architecture. We show that the GTrXL, trained using the same losses, has stability and performance that consistently matches or exceeds a competitive LSTM baseline, including on more reactive tasks where memory is less critical. GTrXL offers an easy-to-train, simple-to-implement but substantially more expressive architectural alternative to the standard multi-layer LSTM ubiquitously used for RL agents in partially-observable environments.
****

A FRAMEWORK FOR ROBUSTNESS CERTIFICATION OF SMOOTHED CLASSIFIERS USING F-DIVERGENCES

Krishnamurthy (Dj) Dvijotham,Jamie Hayes,Borja Balle,Zico Kolter,Chongli Qin,Andras Gyorgy,Kai Xiao,Sven Gowal,Pushmeet Kohli

Formal verification techniques that compute provable guarantees on properties of machine learning models, like robustness to norm-bounded adversarial perturbations, have yielded impressive results. Although most techniques developed so far require knowledge of the architecture of the machine learning model and remain hard to scale to complex prediction pipelines, the method of randomized smoothing has been shown to overcome many of these obstacles. By requiring only black-box access to the underlying model, randomized smoothing scales to large architectures and is agnostic to the internals of the network. However, past work on randomized smoothing has focused on restricted classes of smoothing measures or perturbations (like Gaussian or discrete) and has only been able to prove robustness with respect to simple norm bounds. In this paper we introduce a general framework for proving robustness properties of smoothed machine learning models in the black-box setting. Specifically, we extend randomized smoothing procedures to handle arbitrary smoothing measures and prove robustness of the smoothed classifier by using f-divergences. Our methodology improves upon the state of the art in terms of computation time or certified robustness on several image classification tasks and an audio classification task, with respect to several classes of adversarial perturbations.
****

The Detection of Distributional Discrepancy for Text Generation

Xingyuan Chen,Ping Cai,Peng Jin,Haokun Du,Hongjun Wang,Xinyu Dai,Jiajun Chen

The text generated by neural language models is not as good as the real text. This means that their distributions are different. Generative Adversarial Nets (GAN) are used to alleviate it. However, some researchers argue that GAN variants do not work at all. When both sample quality (such as Bleu) and sample diversity (such as self-Bleu) are taken into account, the GAN variants even are worse than a well-adjusted language model. But, Bleu and self-Bleu can not precisely measure this distributional discrepancy. In fact, how to measure the distributional discrepancy between real text and generated text is still an open problem. In this paper, we theoretically propose two metric functions to measure the distributional difference between real text and generated text. Besides that, a method is put forward to estimate them. First, we evaluate language model with these two functions and find the difference is huge. Then, we try several methods to use the detected discrepancy signal to improve the generator. However the difference becomes even bigger than before. Experimenting on two existing language GANs, the distributional discrepancy between real text and generated text increases with more adversarial learning rounds. It demonstrates both of these language GANs fail.
****

Relative Pixel Prediction For Autoregressive Image Generation

Wang Ling,Chris Dyer,Lei Yu,Lingpeng Kong,Dani Yogatama,Susannah Young

In natural images, transitions between adjacent pixels tend to be smooth and gradual, a fact that has long been exploited in image compression models based on predictive coding. In contrast, existing neural autoregressive image generation models predict the absolute pixel intensities at each position, which is a more challenging problem. In this paper, we propose to predict pixels relatively, by predicting new pixels relative to previously generated pixels (or pixels from the

conditioning context, when available). We show that this form of prediction far e favorably to its absolute counterpart when used independently, but their coord ination under an unified probabilistic model yields optimal performance, as the model learns to predict sharp transitions using the absolute predictor, while ge nerating smooth transitions using the relative predictor.
Experiments on multiple benchmarks for unconditional image generation, image col orization, and super-resolution indicate that our presented mechanism leads to i mprovements in terms of likelihood compared to the absolute prediction counterpa rts.
****

Natural- to formal-language generation using Tensor Product Representations
Kezhen Chen,Qiuyuan Huang,Hamid Palangi,Paul Smolensky,Kenneth D. Forbus,Jianfen g Gao
Generating formal-language represented by relational tuples, such as Lisp progra ms or mathematical expressions, from a natural-language input is an extremely ch allenging task because it requires to explicitly capture discrete symbolic struc tural information from the input to generate the output. Most state-of-the-art n eural sequence models do not explicitly capture such structure information, and thus do not perform well on these tasks. In this paper, we propose a new encoder -decoder model based on Tensor Product Representations (TPRs) for Natural- to Fo rmal-language generation, called TP-N2F. The encoder of TP-N2F employs TPR 'bind ing' to encode natural-language symbolic structure in vector space and the decod er uses TPR 'unbinding' to generate a sequence of relational tuples, each consis ting of a relation (or operation) and a number of arguments, in symbolic space. TP-N2F considerably outperforms LSTM-based Seq2Seq models, creating a new state of the art results on two benchmarks: the MathQA dataset for math problem solvin g, and the AlgoList dataset for program synthesis. Ablation studies show that im provements are mainly attributed to the use of TPRs in both the encoder and deco der to explicitly capture relational structure information for symbolic reasonin g.
****

Three-Head Neural Network Architecture for AlphaZero Learning
Chao Gao,Martin Mueller,Ryan Hayward,Hengshuai Yao,Shangling Jui
The search-based reinforcement learning algorithm AlphaZero has been used as a g eneral method for
mastering two-player games Go, chess and Shogi. One crucial ingredient in AlphaZ ero (and its predecessor AlphaGo Zero) is the two-head network architecture that outputs two estimates --- policy and value --- for one input game state. The me rit of such an architecture is that letting policy and value learning share the same representation substantially improved generalization of the neural net.
A three-head network architecture has been recently proposed that can learn a th ird action-value head on a fixed dataset the same as for two-head net. Also, usi ng the action-value head in Monte Carlo tree search (MCTS) improved the search e fficiency.
However, effectiveness of the three-head network has not been investigated in an AlphaZero style learning paradigm.
In this paper, using the game of Hex as a test domain, we conduct an empirical s tudy of the three-head network architecture in AlpahZero learning. We show that the architecture is also advantageous at the zero-style iterative learning. Spec ifically, we find that three-head network can induce the following benefits: (1) learning can become faster as search takes advantage of the additional action-v alue head; (2) better prediction results than two-head architecture can be achie ved when using additional action-value learning as an auxiliary task.
****

Consistency-Based Semi-Supervised Active Learning: Towards Minimizing Labeling B udget
Mingfei Gao,Zizhao Zhang,Guo Yu,Sercan O. Arik,Larry S. Davis,Tomas Pfister
Active learning (AL) aims to integrate data labeling and model training in a uni fied way, and to minimize the labeling budget by prioritizing the selection of h igh value data that can best improve model performance. Readily-available unlabe

led data are used to evaluate selection mechanisms, but are not used for model t raining in conventional pool-based AL. To minimize the labeling budget, we unify unlabeled sample selection and model training based on two principles. First, w e exploit both labeled and unlabeled data using semi-supervised learning (SSL) t o distill information from unlabeled data that improves representation learning and sample selection. Second, we propose a simple yet effective selection metric that is coherent with the training objective such that the selected samples are effective at improving model performance. Our experimental results demonstrate superior performance with our proposed principles for limited labeled data compa red to alternative AL and SSL combinations. In addition, we study the AL phenom ena of `cold start', which is becoming an increasingly more important factor to enable optimal unification of data labeling, model training and labeling budget minimization. We propose a measure that is found to be empirically correlated w ith the AL target loss. This measure can be used to assist in determining the pr oper start size.

****

Interpretable Network Structure for Modeling Contextual Dependency

Xindian Ma,Peng Zhang,Xiaoliu Mao,Yehua Zhang,Nan Duan,Yuexian Hou,Ming Zhou.

Neural language models have achieved great success in many NLP tasks, to a large extent, due to the ability to capture contextual dependencies among terms in a text. While many efforts have been devoted to empirically explain the connection between the network hyperparameters and the ability to represent the contextual dependency, the theoretical analysis is relatively insufficient. Inspired by th e recent research on the use of tensor space to explain the neural network archi tecture, we explore the interpretable mechanism for neural language models. Spec ifically, we define the concept of separation rank in the language modeling proc ess, in order to theoretically measure the degree of contextual dependencies in a sentence. Then, we show that the lower bound of such a separation rank can rev eal the quantitative relation between the network structure (e.g. depth/width) a nd the modeling ability for the contextual dependency. Especially, increasing th e depth of the neural network can be more effective to improve the ability of mo deling contextual dependency. Therefore, it is important to design an adaptive n etwork to compute the adaptive depth in a task. Inspired by Adaptive Computation Time (ACT), we design an adaptive recurrent network based on the separation ran k to model contextual dependency. Experiments on various NLP tasks have verified the proposed theoretical analysis. We also test our adaptive recurrent neural n etwork in the sentence classification task, and the experiments show that it can achieve better results than the traditional bidirectional LSTM.

****

Policy Tree Network

Zac Wellmer,Sepanta Zeighami,James Kwok

Decision-time planning policies with implicit dynamics models have been shown to work in discrete action spaces with Q learning. However, decision-time planning with implicit dynamics models in continuous action space has proven to be a dif ficult problem. Recent work in Reinforcement Learning has allowed for implicit m odel based approaches to be extended to Policy Gradient methods. In this work we propose Policy Tree Network (PTN). Policy Tree Network lies at the intersection of Model-Based Reinforcement Learning and Model-Free Reinforcement Learning. Po licy Tree Network is a novel approach which, for the first time, demonstrates ho w to leverage an implicit model to perform decision-time planning with Policy Gr adient methods in continuous action spaces. This work is empirically justified o n 8 standard MuJoCo environments so that it can easily be compared with similar work done in this area. Additionally, we offer a lower bound on the worst case c hange in the mean of the policy when tree planning is used and theoretically jus tify our design choices.

****

Padé Activation Units: End-to-end Learning of Flexible Activation Functions in D eep Networks

Alejandro Molina,Patrick Schramowski,Kristian Kersting

The performance of deep network learning strongly depends on the choice of the n

on-linear activation function associated with each neuron. However, deciding on the best activation is non-trivial and the choice depends on the architecture, hyper-parameters, and even on the dataset. Typically these activations are fixed by hand before training. Here, we demonstrate how to eliminate the reliance on first picking fixed activation functions by using flexible parametric rational functions instead. The resulting Padé Activation Units (PAUs) can both approximate common activation functions and also learn new ones while providing compact representations. Our empirical evidence shows that end-to-end learning deep networks with PAUs can increase the predictive performance. Moreover, PAUs pave the way to approximations with provable robustness.
****

Characterize and Transfer Attention in Graph Neural Networks
Mufei Li,Hao Zhang,Xingjian Shi,Minjie Wang,Yixing Guan,Zheng Zhang
Does attention matter and, if so, when and how? Our study on both inductive and transductive learning suggests that datasets have a strong influence on the effects of attention in graph neural networks. Independent of learning setting, task and attention variant, attention mostly degenerate to simple averaging for all three citation networks, whereas they behave strikingly different in the protein-protein interaction networks and molecular graphs: nodes attend to different neighbors per head and get more focused in deeper layers. Consequently, attention distributions become telltale features of the datasets themselves. We further explore the possibility of transferring attention for graph sparsification and show that, when applicable, attention-based sparsification retains enough information to obtain good performance while reducing computational and storage costs. Finally, we point out several possible directions for further study and transfer of attention.
****

Learning to Retrieve Reasoning Paths over Wikipedia Graph for Question Answering
Akari Asai,Kazuma Hashimoto,Hannaneh Hajishirzi,Richard Socher,Caiming Xiong
Answering questions that require multi-hop reasoning at web-scale necessitates retrieving multiple evidence documents, one of which often has little lexical or semantic relationship to the question. This paper introduces a new graph-based recurrent retrieval approach that learns to retrieve reasoning paths over the Wikipedia graph to answer multi-hop open-domain questions. Our retriever model trains a recurrent neural network that learns to sequentially retrieve evidence paragraphs in the reasoning path by conditioning on the previously retrieved documents.
Our reader model ranks the reasoning paths and extracts the answer span included in the best reasoning path.
Experimental results show state-of-the-art results in three open-domain QA datasets, showcasing the effectiveness and robustness of our method. Notably, our method achieves significant improvement in HotpotQA, outperforming the previous best model by more than 14 points.
****

A Baseline for Few-Shot Image Classification
Guneet Singh Dhillon,Pratik Chaudhari,Avinash Ravichandran,Stefano Soatto
Fine-tuning a deep network trained with the standard cross-entropy loss is a strong baseline for few-shot learning. When fine-tuned transductively, this outperforms the current state-of-the-art on standard datasets such as Mini-ImageNet, Tiered-ImageNet, CIFAR-FS and FC-100 with the same hyper-parameters. The simplicity of this approach enables us to demonstrate the first few-shot learning results on the ImageNet-21k dataset. We find that using a large number of meta-training classes results in high few-shot accuracies even for a large number of few-shot classes. We do not advocate our approach as the solution for few-shot learning, but simply use the results to highlight limitations of current benchmarks and few-shot protocols. We perform extensive studies on benchmark datasets to propose a metric that quantifies the "hardness" of a few-shot episode. This metric can be used to report the performance of few-shot algorithms in a more systematic way.
****

Abstract Diagrammatic Reasoning with Multiplex Graph Networks
Duo Wang,Mateja Jamnik,Pietro Lio
Abstract reasoning, particularly in the visual domain, is a complex human abilit y, but it remains a challenging problem for artificial neural learning systems. In this work we propose MXGNet, a multilayer graph neural network for multi-pane l diagrammatic reasoning tasks. MXGNet combines three powerful concepts, namely, object-level representation, graph neural networks and multiplex graphs, for so lving visual reasoning tasks. MXGNet first extracts object-level representations for each element in all panels of the diagrams, and then forms a multi-layer mu ltiplex graph capturing multiple relations between objects across different diag ram panels. MXGNet summarises the multiple graphs extracted from the diagrams of the task, and uses this summarisation to pick the most probable answer from the given candidates. We have tested MXGNet on two types of diagrammatic reasoning tasks, namely Diagram Syllogisms and Raven Progressive Matrices (RPM). For an Eu ler Diagram Syllogism task MXGNet achieves state-of-the-art accuracy of 99.8%. For PGM and RAVEN, two comprehensive datasets for RPM reasoning, MXGNet outperfo rms the state-of-the-art models by a considerable margin.
****
Environmental drivers of systematicity and generalization in a situated agent
Felix Hill,Andrew Lampinen,Rosalia Schneider,Stephen Clark,Matthew Botvinick,Jam es L. McClelland,Adam Santoro
The question of whether deep neural networks are good at generalising beyond the ir immediate training experience is of critical importance for learning-based ap proaches to AI. Here, we consider tests of out-of-sample generalisation that req uire an agent to respond to never-seen-before instructions by manipulating and p ositioning objects in a 3D Unity simulated room. We first describe a comparative ly generic agent architecture that exhibits strong performance on these tests. W e then identify three aspects of the training regime and environment that make a significant difference to its performance: (a) the number of object/word experi ences in the training set; (b) the visual invariances afforded by the agent's pe rspective, or frame of reference; and (c) the variety of visual input inherent i n the perceptual aspect of the agent's perception. Our findings indicate that th e degree of generalisation that networks exhibit can depend critically on partic ulars of the environment in which a given task is instantiated. They further sug gest that the propensity for neural networks to generalise in systematic ways ma y increase if, like human children, those networks have access to many frames of richly varying, multi-modal observations as they learn.
****
SoftAdam: Unifying SGD and Adam for better stochastic gradient descent
Abraham J. Fetterman,Christina H. Kim,Joshua Albrecht
Abstract Stochastic gradient descent (SGD) and Adam are commonly used to optimiz e deep neural networks, but choosing one usually means making tradeoffs between speed, accuracy and stability. Here we present an intuition for why the tradeoff s exist as well as a method for unifying the two in a continuous way. This makes it possible to control the way models are trained in much greater detail. We sh ow that for default parameters, the new algorithm equals or outperforms SGD and Adam across a range of models for image classification tasks and outperforms SGD for language modeling tasks.
****
ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators
Kevin Clark,Minh-Thang Luong,Quoc V. Le,Christopher D. Manning
Masked language modeling (MLM) pre-training methods such as BERT corrupt the inp ut by replacing some tokens with [MASK] and then train a model to reconstruct th e original tokens. While they produce good results when transferred to downstrea m NLP tasks, they generally require large amounts of compute to be effective. As an alternative, we propose a more sample-efficient pre-training task called rep laced token detection. Instead of masking the input, our approach corrupts it by replacing some tokens with plausible alternatives sampled from a small generato r network. Then, instead of training a model that predicts the original identiti es of the corrupted tokens, we train a discriminative model that predicts whethe

r each token in the corrupted input was replaced by a generator sample or not. T horough experiments demonstrate this new pre-training task is more efficient tha n MLM because the task is defined over all input tokens rather than just the sma ll subset that was masked out. As a result, the contextual representations learn ed by our approach substantially outperform the ones learned by BERT given the s ame model size, data, and compute. The gains are particularly strong for small m odels; for example, we train a model on one GPU for 4 days that outperforms GPT (trained using 30x more compute) on the GLUE natural language understanding benc hmark. Our approach also works well at scale, where it performs comparably to Ro BERTa and XLNet while using less than 1/4 of their compute and outperforms them when using the same amount of compute.

****

Evolutionary Population Curriculum for Scaling Multi-Agent Reinforcement Learnin g

Qian Long*,Zihan Zhou*,Abhinav Gupta,Fei Fang,Yi Wu†,Xiaolong Wang†

In multi-agent games, the complexity of the environment can grow exponentially a s the number of agents increases, so it is particularly challenging to learn goo d policies when the agent population is large. In this paper, we introduce Evolu tionary Population Curriculum (EPC), a curriculum learning paradigm that scales up Multi-Agent Reinforcement Learning (MARL) by progressively increasing the pop ulation of training agents in a stage-wise manner. Furthermore, EPC uses an evol utionary approach to fix an objective misalignment issue throughout the curricul um: agents successfully trained in an early stage with a small population are no t necessarily the best candidates for adapting to later stages with scaled popul ations. Concretely, EPC maintains multiple sets of agents in each stage, perform s mix-and-match and fine-tuning over these sets and promotes the sets of agents with the best adaptability to the next stage. We implement EPC on a popular MARL algorithm, MADDPG, and empirically show that our approach consistently outperfo rms baselines by a large margin as the number of agents grows exponentially. The source code and videos can be found at https://sites.google.com/view/epciclr202 0.

****

Amharic Text Normalization with Sequence-to-Sequence Models

Seifedin Shifaw Mohamed,Solomon Teferra Abate (PhD)

All areas of language and speech technology, directly or indirectly, require han dling of real text. In addition to ordinary words and names, the real text conta ins non-standard words (NSWs), including numbers, abbreviations, dates, currency , amounts, and acronyms. Typically, one cannot find NSWs in a dictionary, nor ca n one find their pronunciation by an application of ordinary letter-to-sound rul es. It is desirable to normalize text by replacing such non-standard words with a consistently formatted and contextually appropriate variant in several NLP app lications. To address this challenge, in this paper, we model the problem as cha racter-level sequence-to-sequence learning where we map a sequence of input char acters to a sequence of output words. It consists of two neural networks, the en coder network, and the decoder network. The encoder maps the input characters to a fixed dimensional vector and the decoder generates the output words. We have achieved an accuracy of 94.8 % which is promising given the resource we use.

****

Thinking While Moving: Deep Reinforcement Learning with Concurrent Control

Ted Xiao,Eric Jang,Dmitry Kalashnikov,Sergey Levine,Julian Ibarz,Karol Hausman,A lexander Herzog

We study reinforcement learning in settings where sampling an action from the po licy must be done concurrently with the time evolution of the controlled system, such as when a robot must decide on the next action while still performing the previous action. Much like a person or an animal, the robot must think and move at the same time, deciding on its next action before the previous one has comple ted. In order to develop an algorithmic framework for such concurrent control pr oblems, we start with a continuous-time formulation of the Bellman equations, an d then discretize them in a way that is aware of system delays. We instantiate t

his new class of approximate dynamic programming methods via a simple architectu
ral extension to existing value-based deep reinforcement learning algorithms. We
 evaluate our methods on simulated benchmark tasks and a large-scale robotic gra
sping task where the robot must "think while moving."
****

RATE-DISTORTION OPTIMIZATION GUIDED AUTOENCODER FOR GENERATIVE APPROACH
Keizo Kato,Jing Zhou,Akira Nakagawa
In the generative model approach of machine learning, it is essential to acquire
 an accurate probabilistic model and compress the dimension of data for easy tre
atment. However, in the conventional deep-autoencoder based generative model suc
h as VAE, the probability of the real space cannot be obtained correctly from th
at of in the latent space, because the scaling between both spaces is not contro
lled. This has also been an obstacle to quantifying the impact of the variation
 of latent variables on data. In this paper, we propose a method to learn paramet
ric probability distribution and autoencoder simultaneously based on Rate-Distor
tion Optimization to support scaling control. It is proved theoretically and exp
erimentally that (i) the probability distribution of the latent space obtained b
y this model is proportional to the probability distribution of the real space b
ecause Jacobian between two spaces is constant: (ii) our model behaves as non-li
near PCA, which enables to evaluate the influence of latent variables on data. F
urthermore, to verify the usefulness on the practical application, we evaluate i
ts performance in unsupervised anomaly detection and outperform current state-of
-the-art methods.
****

On the expected running time of nonconvex optimization with early stopping
Thomas Flynn,Kwang Min Yu,Abid Malik,Shinjae Yoo,Nicholas D'Imperio
This work examines the convergence of stochastic gradient algorithms that use ea
rly stopping based on a validation function, wherein optimization ends when the
 magnitude of a validation function gradient drops below a threshold. We derive c
onditions that guarantee this stopping rule is well-defined and analyze the expe
cted number of iterations and gradient evaluations needed to meet this criteria.
 The guarantee accounts for the distance between the training and validation set
s, measured with the Wasserstein distance. We develop the approach for stochasti
c gradient descent (SGD), allowing for biased update directions subject to a Lya
punov condition. We apply the approach to obtain new bounds on the expected runn
ing time of several algorithms, including Decentralized SGD (DSGD), a variant of
 decentralized SGD, known as \textit{Stacked SGD}, and the stochastic variance r
educed gradient (SVRG) algorithm. Finally, we consider the generalization proper
ties of the iterate returned by early stopping.
****

Multiagent Reinforcement Learning in Games with an Iterated Dominance Solution
Yoram Bachrach,Tor Lattimore,Marta Garnelo,Julien Perolat,David Balduzzi,Thomas
Anthony,Satinder Singh,Thore Graepel
Multiagent reinforcement learning (MARL) attempts to optimize policies of intell
igent agents interacting in the same environment. However, it may fail to conver
ge to a Nash equilibrium in some games. We study independent MARL under the mor
e demanding solution concept of iterated elimination of strictly dominated strat
egies. In dominance solvable games, if players iteratively eliminate strictly d
ominated strategies until no further strategies can be eliminated, we obtain a s
ingle strategy profile. We show that convergence to the iterated dominance solut
ion is guaranteed for several reinforcement learning algorithms (for multiple in
dependent learners). We illustrate an application of our results by studying mec
hanism design for principal-agent problems, where a principal wishes to incentiv
ize agents to exert costly effort in a joint project when it can only observe wh
ether the project succeeded, but not whether agents actually exerted effort. We
 show that MARL converges to the desired outcome if the rewards are designed so t
hat exerting effort is the iterated dominance solution, but fails if it is merel
y a Nash equilibrium.
****

CP-GAN: Towards a Better Global Landscape of GANs

Ruoyu Sun,Tiantian Fang,Alex Schwing
GANs have been very popular in data generation and unsupervised learning, but our understanding of GAN training is still very limited. One major reason is that GANs are often formulated as non-convex-concave min-max optimization. As a result, most recent studies focused on the analysis in the local region around the equilibrium. In this work, we perform a global analysis of GANs from two perspectives: the global landscape of the outer-optimization problem and the global behavior of the gradient descent dynamics. We find that the original GAN has exponentially many bad strict local minima which are perceived as mode-collapse, and the training dynamics (with linear discriminators) cannot escape mode collapse. To address these issues, we propose a simple modification to the original GAN, by coupling the generated samples and the true samples. We prove that the new formulation has no bad basins, and its training dynamics (with linear discriminators) has a Lyapunov function that leads to global convergence. Our experiments on standard datasets show that this simple loss outperforms the original GAN and WGAN-GP.
****

Jacobian Adversarially Regularized Networks for Robustness
Alvin Chan,Yi Tay,Yew Soon Ong,Jie Fu
Adversarial examples are crafted with imperceptible perturbations with the intent to fool neural networks. Against such attacks, adversarial training and its variants stand as the strongest defense to date. Previous studies have pointed out that robust models that have undergone adversarial training tend to produce more salient and interpretable Jacobian matrices than their non-robust counterparts. A natural question is whether a model trained with an objective to produce salient Jacobian can result in better robustness. This paper answers this question with affirmative empirical results. We propose Jacobian Adversarially Regularized Networks (JARN) as a method to optimize the saliency of a classifier's Jacobian by adversarially regularizing the model's Jacobian to resemble natural training images. Image classifiers trained with JARN show improved robust accuracy compared to standard models on the MNIST, SVHN and CIFAR-10 datasets, uncovering a new angle to boost robustness without using adversarial training.
****

Gradient Descent can Learn Less Over-parameterized Two-layer Neural Networks on Classification Problems
Atsushi Nitanda,Geoffrey Chinot,Taiji Suzuki
Recently, several studies have proven the global convergence and generalization abilities of the gradient descent method for two-layer ReLU networks. Most studies especially focused on the regression problems with the squared loss function, except for a few, and the importance of the positivity of the neural tangent kernel has been pointed out. However, the performance of gradient descent on classification problems using the logistic loss function has not been well studied, and further investigation of this problem structure is possible. In this work, we demonstrate that the separability assumption using a neural tangent model is more reasonable than the positivity condition of the neural tangent kernel and provide a refined convergence analysis of the gradient descent for two-layer networks with smooth activations. A remarkable point of our result is that our convergence and generalization bounds have much better dependence on the network width in comparison to related studies. Consequently, our theory significantly enlarges a class of over-parameterized networks with provable generalization ability, with respect to the network width, while most studies require much higher over-parameterization.
****

Improving Federated Learning Personalization via Model Agnostic Meta Learning
Yihan Jiang,Jakub Kone■ný,Keith Rush,Sreeram Kannan
Federated Learning (FL) refers to learning a high quality global model based on decentralized data storage, without ever copying the raw data. A natural scenario arises with data created on mobile phones by the activity of their users. Given the typical data heterogeneity in such situations, it is natural to ask how can the global model be personalized for every such device, individually. In this

work, we point out that the setting of Model Agnostic Meta Learning (MAML), where one optimizes for a fast, gradient-based, few-shot adaptation to a heterogeneous distribution of tasks, has a number of similarities with the objective of personalization for FL. We present FL as a natural source of practical applications for MAML algorithms, and make the following observations. 1) The popular FL algorithm, Federated Averaging, can be interpreted as a meta learning algorithm. 2) Careful fine-tuning can yield a global model with higher accuracy, which is at the same time easier to personalize. However, solely optimizing for the global model accuracy yields a weaker personalization result. 3) A model trained using a standard datacenter optimization method is much harder to personalize, compared to one trained using Federated Averaging, supporting the first claim. These results raise new questions for FL, MAML, and broader ML research.

****

Towards Verified Robustness under Text Deletion Interventions

Johannes Welbl,Po-Sen Huang,Robert Stanforth,Sven Gowal,Krishnamurthy (Dj) Dvijotham,Martin Szummer,Pushmeet Kohli

Neural networks are widely used in Natural Language Processing, yet despite their empirical successes, their behaviour is brittle: they are both over-sensitive to small input changes, and under-sensitive to deletions of large fractions of input text. This paper aims to tackle under-sensitivity in the context of natural language inference by ensuring that models do not become more confident in their predictions as arbitrary subsets of words from the input text are deleted. We develop a novel technique for formal verification of this specification for models based on the popular decomposable attention mechanism by employing the efficient yet effective interval bound propagation (IBP) approach. Using this method we can efficiently prove, given a model, whether a particular sample is free from the under-sensitivity problem. We compare different training methods to address under-sensitivity, and compare metrics to measure it. In our experiments on the SNLI and MNLI datasets, we observe that IBP training leads to a significantly improved verified accuracy. On the SNLI test set, we can verify 18.4% of samples, a substantial improvement over only 2.8% using standard training.

****

Discovering Topics With Neural Topic Models Built From PLSA Loss

sileye ba

In this paper we present a model for unsupervised topic discovery in texts corpora. The proposed model uses documents, words, and topics lookup table embedding as neural network model parameters to build probabilities of words given topics, and probabilities of topics given documents. These probabilities are used to recover by marginalization probabilities of words given documents. For very large corpora where the number of documents can be in the order of billions, using a neural auto-encoder based document embedding is more scalable then using a lookup table embedding as classically done. We thus extended the lookup based document embedding model to continuous auto-encoder based model. Our models are trained using probabilistic latent semantic analysis (PLSA) assumptions. We evaluated our models on six datasets with a rich variety of contents. Conducted experiments demonstrate that the proposed neural topic models are very effective in capturing relevant topics. Furthermore, considering perplexity metric, conducted evaluation benchmarks show that our topic models outperform latent Dirichlet allocation (LDA) model which is classically used to address topic discovery tasks.

****

And the Bit Goes Down: Revisiting the Quantization of Neural Networks

Pierre Stock,Armand Joulin,Rémi Gribonval,Benjamin Graham,Hervé Jégou

In this paper, we address the problem of reducing the memory footprint of convolutional network architectures. We introduce a vector quantization method that aims at preserving the quality of the reconstruction of the network outputs rather than its weights. The principle of our approach is that it minimizes the loss reconstruction error for in-domain inputs. Our method only requires a set of unlabelled data at quantization time and allows for efficient inference on CPU by using byte-aligned codebooks to store the compressed weights. We validate our approach by quantizing a high performing ResNet-50 model to a memory size of 5MB (20

x compression factor) while preserving a top-1 accuracy of 76.1% on ImageNet obj
ect classification and by compressing a Mask R-CNN with a 26x factor.
****

Meta-Learning Runge-Kutta
Nadine Behrmann,Patrick Schramowski,Kristian Kersting
Initial value problems, i.e. differential equations with specific, initial condi
tions, represent a classic problem within the field of ordinary differential equ
ations(ODEs). While the simplest types of ODEs may have closed-form solutions, m
ost interesting cases typically rely on iterative schemes for numerical integrat
ion such as the family of Runge-Kutta methods. They are, however, sensitive to t
he strategy the step size is adapted during integration, which has to be chosen
by the experimenter. In this paper, we show how the design of a step size contro
ller can be cast as a learning problem, allowing deep networks to learn to explo
it structure in the initial value problem at hand in an automatic way. The key i
ngredients for the resulting Meta-Learning Runge-Kutta (MLRK) are the developmen
t of a good performance measure and the identification of suitable input feature
s. Traditional approaches suggest the local error estimates as input to the cont
roller. However, by studying the characteristics of the local error function we
show that including the partial derivatives of the initial value problem is favo
rable. Our experiments demonstrate considerable benefits over traditional approa
ches. In particular, MLRK is able to mitigate sudden spikes in the local error f
unction by a faster adaptation of the step size. More importantly, the additiona
l information in the form of partial derivatives and function values leads to a
substantial improvement in performance. The source code can be found at https://
www.dropbox.com/sh/rkctdfhkosywnnx/AABKadysCR8-aHW_0kb6vCtSa?dl=0
****

RGBD-GAN: Unsupervised 3D Representation Learning From Natural Image Datasets vi
a RGBD Image Synthesis
Atsuhiro Noguchi,Tatsuya Harada
Understanding three-dimensional (3D) geometries from two-dimensional (2D) images
 without any labeled information is promising for understanding the real world w
ithout incurring annotation cost. We herein propose a novel generative model, RG
BD-GAN, which achieves unsupervised 3D representation learning from 2D images. T
he proposed method enables camera parameter--conditional image generation and de
pth image generation without any 3D annotations, such as camera poses or depth.
We use an explicit 3D consistency loss for two RGBD images generated from differ
ent camera parameters, in addition to the ordinal GAN objective. The loss is sim
ple yet effective for any type of image generator such as DCGAN and StyleGAN to
be conditioned on camera parameters. Through experiments, we demonstrated that t
he proposed method could learn 3D representations from 2D images with various ge
nerator architectures.
****

Provable Benefit of Orthogonal Initialization in Optimizing Deep Linear Networks
Wei Hu,Lechao Xiao,Jeffrey Pennington
The selection of initial parameter values for gradient-based optimization of dee
p neural networks is one of the most impactful hyperparameter choices in deep le
arning systems, affecting both convergence times and model performance. Yet desp
ite significant empirical and theoretical analysis, relatively little has been p
roved about the concrete effects of different initialization schemes. In this wo
rk, we analyze the effect of initialization in deep linear networks, and provide
 for the first time a rigorous proof that drawing the initial weights from the o
rthogonal group speeds up convergence relative to the standard Gaussian initiali
zation with iid weights. We show that for deep networks, the width needed for ef
ficient convergence to a global minimum with orthogonal initializations is indep
endent of the depth, whereas the width needed for efficient convergence with Gau
ssian initializations scales linearly in the depth. Our results demonstrate how
the benefits of a good initialization can persist throughout learning, suggestin
g an explanation for the recent empirical successes found by initializing very d
eep non-linear networks according to the principle of dynamical isometry.
****

Hallucinative Topological Memory for Zero-Shot Visual Planning
Kara Liu,Thanard Kurutach,Pieter Abbeel,Aviv Tamar
In visual planning (VP), an agent learns to plan goal-directed behavior from observations of a dynamical system obtained offline, e.g., images obtained from self-supervised robot interaction. VP algorithms essentially combine data-driven perception and planning, and are important for robotic manipulation and navigation domains, among others. A recent and promising approach to VP is the semi-parametric topological memory (SPTM) method, where image samples are treated as nodes in a graph, and the connectivity in the graph is learned using deep image classification. Thus, the learned graph represents the topological connectivity of the data, and planning can be performed using conventional graph search methods. However, training SPTM necessitates a suitable loss function for the connectivity classifier, which requires non-trivial manual tuning. More importantly, SPTM is constricted in its ability to generalize to changes in the domain, as its graph is constructed from direct observations and thus requires collecting new samples for planning. In this paper, we propose Hallucinative Topological Memory (HTM), which overcomes these shortcomings. In HTM, instead of training a discriminative classifier we train an energy function using contrastive predictive coding. In addition, we learn a conditional VAE model that generates samples given a context image of the domain, and use these hallucinated samples for building the connectivity graph, allowing for zero-shot generalization to domain changes. In simulated domains, HTM outperforms conventional SPTM and visual foresight methods in terms of both plan quality and success in long-horizon planning.
****
Learning Good Policies By Learning Good Perceptual Models
Yilun Du,Phillip Isola
Reinforcement learning (RL) has led to increasingly complex looking behavior in recent years. However, such complexity can be misleading and hides over-fitting. We find that visual representations may be a useful metric of complexity, and both correlates well objective optimization and causally effects reward optimization. We then propose curious representation learning (CRL) which allows us to use better visual representation learning algorithms to correspondingly increase visual representation in policy through an intrinsic objective on both simulated environments and transfer to real images. Finally, we show better visual representations induced by CRL allows us to obtain better performance on Atari without any reward than other curiosity objectives.
****
Implementation Matters in Deep RL: A Case Study on PPO and TRPO
Logan Engstrom,Andrew Ilyas,Shibani Santurkar,Dimitris Tsipras,Firdaus Janoos,Larry Rudolph,Aleksander Madry
We study the roots of algorithmic progress in deep policy gradient algorithms through a case study on two popular algorithms: Proximal Policy Optimization (PPO) and Trust Region Policy Optimization (TRPO). Specifically, we investigate the consequences of "code-level optimizations:" algorithm augmentations found only in implementations or described as auxiliary details to the core algorithm. Seemingly of secondary importance, such optimizations turn out to have a major impact on agent behavior. Our results show that they (a) are responsible for most of PPO's gain in cumulative reward over TRPO, and (b) fundamentally change how RL methods function. These insights show the difficulty, and importance, of attributing performance gains in deep reinforcement learning.

****
A Closer Look at Deep Policy Gradients
Andrew Ilyas,Logan Engstrom,Shibani Santurkar,Dimitris Tsipras,Firdaus Janoos,Larry Rudolph,Aleksander Madry
We study how the behavior of deep policy gradient algorithms reflects the conceptual framework motivating their development. To this end, we propose a fine-grained analysis of state-of-the-art methods based on key elements of this framework: gradient estimation, value prediction, and optimization landscapes. Our results show that the behavior of deep policy gradient algorithms often deviates f

rom what their motivating framework would predict: surrogate rewards do not matc h the true reward landscape, learned value estimators fail to fit the true value function, and gradient estimates poorly correlate with the "true" gradient. The mismatch between predicted and empirical behavior we uncover highlights our poo r understanding of current methods, and indicates the need to move beyond curren t benchmark-centric evaluation methods.
****

Plug and Play Language Models: A Simple Approach to Controlled Text Generation
Sumanth Dathathri,Andrea Madotto,Janice Lan,Jane Hung,Eric Frank,Piero Molino,Ja son Yosinski,Rosanne Liu
Large transformer-based language models (LMs) trained on huge text corpora have shown unparalleled generation capabilities. However, controlling attributes of t he generated language (e.g. switching topic or sentiment) is difficult without m odifying the model architecture or fine-tuning on attribute-specific data and en tailing the significant cost of retraining. We propose a simple alternative: the Plug and Play Language Model (PPLM) for controllable language generation, which combines a pretrained LM with one or more simple attribute classifiers that gui de text generation without any further training of the LM. In the canonical scen ario we present, the attribute models are simple classifiers consisting of a use r-specified bag of words or a single learned layer with 100,000 times fewer para meters than the LM. Sampling entails a forward and backward pass in which gradie nts from the attribute model push the LM's hidden activations and thus guide the generation. Model samples demonstrate control over a range of topics and sentim ent styles, and extensive automated and human annotated evaluations show attribu te alignment and fluency. PPLMs are flexible in that any combination of differen tiable attribute models may be used to steer text generation, which will allow f or diverse and creative applications beyond the examples given in this paper.
****

Efficient High-Dimensional Data Representation Learning via Semi-Stochastic Bloc k Coordinate Descent Methods
Bingkun Wei,Yangyang Li,Fanhua Shang,Yuanyuan Liu,Hongying Liu,Shengmei Shen
With the increase of data volume and data dimension, sparse representation learn ing attracts more and more attention. For high-dimensional data, randomized bloc k coordinate descent methods perform well because they do not need to calculate the gradient along the whole dimension. Existing hard thresholding algorithms ev aluate gradients followed by a hard thresholding operation to update the model p arameter, which leads to slow convergence. To address this issue, we propose a n ovel hard thresholding algorithm, called Semi-stochastic Block Coordinate Descen t Hard Thresholding Pursuit (SBCD-HTP). Moreover, we present its sparse and asyn chronous parallel variants. We theoretically analyze the convergence properties of our algorithms, which show that they have a significantly lower hard threshol ding complexity than existing algorithms. Our empirical evaluations on real-worl d datasets and face recognition tasks demonstrate the superior performance of ou r algorithms for sparsity-constrained optimization problems.
****

Understanding and Robustifying Differentiable Architecture Search
Arber Zela,Thomas Elsken,Tonmoy Saikia,Yassine Marrakchi,Thomas Brox,Frank Hutte r
Differentiable Architecture Search (DARTS) has attracted a lot of attention due to its simplicity and small search costs achieved by a continuous relaxation and an approximation of the resulting bi-level optimization problem. However, DART S does not work robustly for new problems: we identify a wide range of search sp aces for which DARTS yields degenerate architectures with very poor test perform ance. We study this failure mode and show that, while DARTS successfully minimiz es validation loss, the found solutions generalize poorly when they coincide wit h high validation loss curvature in the architecture space. We show that by add ing one of various types of regularization we can robustify DARTS to find soluti ons with less curvature and better generalization properties. Based on these obs ervations, we propose several simple variations of DARTS that perform substantia lly more robustly in practice. Our observations are robust across five search s

paces on three image classification tasks and also hold for the very different d
omains of disparity estimation (a dense regression task) and language modelling.
****

Rethinking the Hyperparameters for Fine-tuning
Hao Li,Pratik Chaudhari,Hao Yang,Michael Lam,Avinash Ravichandran,Rahul Bhotika,
Stefano Soatto
Fine-tuning from pre-trained ImageNet models has become the de-facto standard fo
r various computer vision tasks. Current practices for fine-tuning typically inv
olve selecting an ad-hoc choice of hyperparameters and keeping them fixed to val
ues normally used for training from scratch. This paper re-examines several comm
on practices of setting hyperparameters for fine-tuning. Our findings are based
 on extensive empirical evaluation for fine-tuning on various transfer learning
benchmarks. (1) While prior works have thoroughly investigated learning rate and
 batch size, momentum for fine-tuning is a relatively unexplored parameter. We f
ind that the value of momentum also affects fine-tuning performance and connect
it with previous theoretical findings. (2) Optimal hyperparameters for fine-tun
ing, in particular, the effective learning rate, are not only dataset dependent
but also sensitive to the similarity between the source domain and target domain
. This is in contrast to hyperparameters for training from scratch. (3) Referenc
e-based regularization that keeps models close to the initial model does not nec
essarily apply for "dissimilar" datasets. Our findings challenge common practice
s of fine-tuning and encourages deep learning practitioners to rethink the hyper
parameters for fine-tuning.
****

UNITER: Learning UNiversal Image-TExt Representations
Yen-Chun Chen,Linjie Li,Licheng Yu,Ahmed El Kholy,Faisal Ahmed,Zhe Gan,Yu Cheng,
Jingjing Liu
Joint image-text embedding is the bedrock for most Vision-and-Language (V+L) tas
ks, where multimodality inputs are jointly processed for visual and textual unde
rstanding. In this paper, we introduce UNITER, a UNiversal Image-TExt Representa
tion, learned through large-scale pre-training over four image-text datasets (CO
CO, Visual Genome, Conceptual Captions, and SBU Captions), which can power heter
ogeneous downstream V+L tasks with joint multimodal embeddings. We design three
pre-training tasks: Masked Language Modeling (MLM), Image-Text Matching (ITM), a
nd Masked Region Modeling (MRM, with three variants). Different from concurrent
work on multimodal pre-training that apply joint random masking to both modaliti
es, we use Conditioned Masking on pre-training tasks (i.e., masked language/regi
on modeling is conditioned on full observation of image/text). Comprehensive ana
lysis shows that conditioned masking yields better performance than unconditione
d masking. We also conduct a thorough ablation study to find an optimal combinat
ion of pre-training tasks for UNITER. Extensive experiments show that UNITER ach
ieves new state of the art across six V+L tasks over nine datasets, including Vi
sual Question Answering, Image-Text Retrieval, Referring Expression Comprehensio
n, Visual Commonsense Reasoning, Visual Entailment, and NLVR2.
****

Self-Supervised GAN Compression
Chong Yu,Jeff Pool
Deep learning's success has led to larger and larger models to handle more and m
ore complex tasks; trained models can contain millions of parameters. These larg
e models are compute- and memory-intensive, which makes it a challenge to deploy
 them with minimized latency, throughput, and storage requirements. Some model c
ompression methods have been successfully applied on image classification and de
tection or language models, but there has been very little work compressing gene
rative adversarial networks (GANs) performing complex tasks. In this paper, we s
how that a standard model compression technique, weight pruning, cannot be appli
ed to GANs using existing methods. We then develop a self-supervised compression
 technique which uses the trained discriminator to supervise the training of a c
ompressed generator. We show that this framework has a compelling performance to
 high degrees of sparsity, generalizes well to new tasks and models, and enables
 meaningful comparisons between different pruning granularities.

****
Retrieving Signals in the Frequency Domain with Deep Complex Extractors
Chiheb Trabelsi,Olexa Bilaniuk,Ousmane Dia,Ying Zhang,Mirco Ravanelli,Jonathan Binas,Negar Rostamzadeh,Christopher J Pal
Recent advances have made it possible to create deep complex-valued neural networks. Despite this progress, the potential power of fully complex intermediate computations and representations has not yet been explored for many challenging learning problems. Building on recent advances, we propose a novel mechanism for extracting signals in the frequency domain. As a case study, we perform audio source separation in the Fourier domain. Our extraction mechanism could be regarded as a local ensembling method that combines a complex-valued convolutional version of Feature-Wise Linear Modulation (FiLM) and a signal averaging operation. We also introduce a new explicit amplitude and phase-aware loss, which is scale and time invariant, taking into account the complex-valued components of the spectrogram. Using the Wall Street Journal Dataset, we compare our phase-aware loss to several others that operate both in the time and frequency domains and demonstrate the effectiveness of our proposed signal extraction method and proposed loss. When operating in the complex-valued frequency domain, our deep complex-valued network substantially outperforms its real-valued counterparts even with half the depth and a third of the parameters. Our proposed mechanism improves significantly deep complex-valued networks' performance and we demonstrate the usefulness of its regularizing effect.
****
Query2box: Reasoning over Knowledge Graphs in Vector Space Using Box Embeddings
Hongyu Ren*,Weihua Hu*,Jure Leskovec
Answering complex logical queries on large-scale incomplete knowledge graphs (KGs) is a fundamental yet challenging task. Recently, a promising approach to this problem has been to embed KG entities as well as the query into a vector space such that entities that answer the query are embedded close to the query. However, prior work models queries as single points in the vector space, which is problematic because a complex query represents a potentially large set of its answer entities, but it is unclear how such a set can be represented as a single point. Furthermore, prior work can only handle queries that use conjunctions ($\wedge$) and existential quantifiers ($\exists$). Handling queries with logical disjunctions ($\vee$) remains an open problem. Here we propose query2box, an embedding-based framework for reasoning over arbitrary queries with $\wedge$, $\vee$, and $\exists$ operators in massive and incomplete KGs. Our main insight is that queries can be embedded as boxes (i.e., hyper-rectangles), where a set of points inside the box corresponds to a set of answer entities of the query. We show that conjunctions can be naturally represented as intersections of boxes and also prove a negative result that handling disjunctions would require embedding with dimension proportional to the number of KG entities. However, we show that by transforming queries into a Disjunctive Normal Form, query2box is capable of handling arbitrary logical queries with $\wedge$, $\vee$, $\exists$ in a scalable manner. We demonstrate the effectiveness of query2box on two large KGs and show that query2box achieves up to 25% relative improvement over the state of the art.

****
Implementing Inductive bias for different navigation tasks through diverse RNN attrractors
Tie XU,Omri Barak
Navigation is crucial for animal behavior and is assumed to require an internal representation of the external environment, termed a cognitive map. The precise form of this representation is often considered to be a metric representation of space. An internal representation, however, is judged by its contribution to performance on a given task, and may thus vary between different types of navigation tasks. Here we train a recurrent neural network that controls an agent performing several navigation tasks in a simple environment. To focus on internal representations, we split learning into a task-agnostic pre-training stage that modifies internal connectivity and a task-specific Q learning stage that controls th

e network's output. We show that pre-training shapes the attractor landscape of the networks, leading to either a continuous attractor, discrete attractors or a disordered state. These structures induce bias onto the Q-Learning phase, leading to a performance pattern across the tasks corresponding to metric and topological regularities. Our results show that, in recurrent networks, inductive bias takes the form of attractor landscapes -- which can be shaped by pre-training and analyzed using dynamical systems methods. Furthermore, we demonstrate that non-metric representations are useful for navigation tasks.
****

Disentangling Style and Content in Anime Illustrations
Sitao Xiang,Hao Li
Existing methods for AI-generated artworks still struggle with generating high-quality stylized content, where high-level semantics are preserved, or separating fine-grained styles from various artists. We propose a novel Generative Adversarial Disentanglement Network which can disentangle two complementary factors of variations when only one of them is labelled in general, and fully decompose complex anime illustrations into style and content in particular. Training such model is challenging, since given a style, various content data may exist but not the other way round. Our approach is divided into two stages, one that encodes an input image into a style independent content, and one based on a dual-conditional generator. We demonstrate the ability to generate high-fidelity anime portraits with a fixed content and a large variety of styles from over a thousand artists, and vice versa, using a single end-to-end network and with applications in style transfer. We show this unique capability as well as superior output to the current state-of-the-art.
****

Dynamic Instance Hardness
Tianyi Zhou,Shengjie Wang,Jeff A. Bilmes
We introduce dynamic instance hardness (DIH) to facilitate the training of machine learning models. DIH is a property of each training sample and is computed as the running mean of the sample's instantaneous hardness as measured over the training history. We use DIH to evaluate how well a model retains knowledge about each training sample over time. We find that for deep neural nets (DNNs), the DIH of a sample in relatively early training stages reflects its DIH in later stages and as a result, DIH can be effectively used to reduce the set of training samples in future epochs. Specifically, during each epoch, only samples with high DIH are trained (since they are historically hard) while samples with low DIH can be safely ignored. DIH is updated each epoch only for the selected samples, so it does not require additional computation. Hence, using DIH during training leads to an appreciable speedup. Also, since the model is focused on the historically more challenging samples, resultant models are more accurate. The above, when formulated as an algorithm, can be seen as a form of curriculum learning, so we call our framework DIH curriculum learning (or DIHCL). The advantages of DIHCL, compared to other curriculum learning approaches, are: (1) DIHCL does not require additional inference steps over the data not selected by DIHCL in each epoch, (2) the dynamic instance hardness, compared to static instance hardness (e.g., instantaneous loss), is more stable as it integrates information over the entire training history up to the present time. Making certain mathematical assumptions, we formulate the problem of DIHCL as finding a curriculum that maximizes a multi-set function $f(\cdot)$, and derive an approximation bound for a DIH-produced curriculum relative to the optimal curriculum. Empirically, DIHCL-trained DNNs significantly outperform random mini-batch SGD and other recently developed curriculum learning methods in terms of efficiency, early-stage convergence, and final performance, and this is shown in training several state-of-the-art DNNs on 11 modern datasets.
****

Multi-step Greedy Policies in Model-Free Deep Reinforcement Learning
Yonathan Efroni,Manan Tomar,Mohammad Ghavamzadeh
Multi-step greedy policies have been extensively used in model-based Reinforcement Learning (RL) and in the case when a model of the environment is available (e

.g., in the game of Go). In this work, we explore the benefits of multi-step greedy policies in model-free RL when employed in the framework of multi-step Dynamic Programming (DP): multi-step Policy and Value Iteration. These algorithms iteratively solve short-horizon decision problems and converge to the optimal solution of the original one. By using model-free algorithms as solvers of the short-horizon problems we derive fully model-free algorithms which are instances of the multi-step DP framework. As model-free algorithms are prone to instabilities w.r.t. the decision problem horizon, this simple approach can help in mitigating these instabilities and results in an improved model-free algorithms. We test this approach and show results on both discrete and continuous control problems.
****

A Random Matrix Perspective on Mixtures of Nonlinearities in High Dimensions
Ben Adlam,Jake Levinson,Jeffrey Pennington
One of the distinguishing characteristics of modern deep learning systems is that they typically employ neural network architectures that utilize enormous numbers of parameters, often in the millions and sometimes even in the billions. While this paradigm has inspired significant research on the properties of large networks, relatively little work has been devoted to the fact that these networks are often used to model large complex datasets, which may themselves contain millions or even billions of constraints. In this work, we focus on this high-dimensional regime in which both the dataset size and the number of features tend to infinity. We analyze the performance of a simple regression model trained on the random features $F=f(WX+B)$ for a random weight matrix W and random bias vector B, obtaining an exact formula for the asymptotic training error on a noisy autoencoding task. The role of the bias can be understood as parameterizing a distribution over activation functions, and our analysis actually extends to general such distributions, even those not expressible with a traditional additive bias. Intruigingly, we find that a mixture of nonlinearities can outperform the best single nonlinearity on the noisy autoecndoing task, suggesting that mixtures of nonlinearities might be useful for approximate kernel methods or neural network architecture design.
****

LIA: Latently Invertible Autoencoder with Adversarial Learning
Jiapeng Zhu,Deli Zhao,Bolei Zhou,Bo Zhang
Deep generative models such as Variational AutoEncoder (VAE) and Generative Adversarial Network (GAN) play an increasingly important role in machine learning and computer vision. However, there are two fundamental issues hindering their real-world applications: the difficulty of conducting variational inference in VAE and the functional absence of encoding real-world samples in GAN. In this paper, we propose a novel algorithm named Latently Invertible Autoencoder (LIA) to address the above two issues in one framework. An invertible network and its inverse mapping are symmetrically embedded in the latent space of VAE. Thus the partial encoder first transforms the input into feature vectors and then the distribution of these feature vectors is reshaped to fit a prior by the invertible network. The decoder proceeds in the reverse order of the encoder's composite mappings. A two-stage stochasticity-free training scheme is designed to train LIA via adversarial learning, in the sense that the decoder of LIA is first trained as a standard GAN with the invertible network and then the partial encoder is learned from an autoencoder by detaching the invertible network from LIA. Experiments conducted on the FFHQ face dataset and three LSUN datasets validate the effectiveness of LIA for inference and generation.
****

PCMC-Net: Feature-based Pairwise Choice Markov Chains
Alix Lhéritier
Pairwise Choice Markov Chains (PCMC) have been recently introduced to overcome limitations of choice models based on traditional axioms unable to express empirical observations from modern behavior economics like context effects occurring when a choice between two options is altered by adding a third alternative. The inference approach that estimates the transition rates between each possible pair of alternatives via maximum likelihood suffers when the examples of each altern

ative are scarce and is inappropriate when new alternatives can be observed at t
est time. In this work, we propose an amortized inference approach for PCMC by e
mbedding its definition into a neural network that represents transition rates a
s a function of the alternatives' and individual's features. We apply our constr
uction to the complex case of airline itinerary booking where singletons are com
mon (due to varying prices and individual-specific itineraries), and context eff
ects and behaviors strongly dependent on market segments are observed. Experimen
ts show our network significantly outperforming, in terms of prediction accuracy
 and logarithmic loss, feature engineered standard and latent class Multinomial
Logit models as well as recent machine learning approaches.
****

Multi-Agent Interactions Modeling with Correlated Policies
Minghuan Liu,Ming Zhou,Weinan Zhang,Yuzheng Zhuang,Jun Wang,Wulong Liu,Yong Yu
In multi-agent systems, complex interacting behaviors arise due to the high corr
elations among agents. However, previous work on modeling multi-agent interactio
ns from demonstrations is primarily constrained by assuming the independence amo
ng policies and their reward structures.
In this paper, we cast the multi-agent interactions modeling problem into a mult
i-agent imitation learning framework with explicit modeling of correlated polici
es by approximating opponents’ policies, which can recover agents' policies that
 can regenerate similar interactions. Consequently, we develop a Decentralized A
dversarial Imitation Learning algorithm with Correlated policies (CoDAIL), which
 allows for decentralized training and execution. Various experiments demonstrat
e that CoDAIL can better regenerate complex interactions close to the demonstrat
ors and outperforms state-of-the-art multi-agent imitation learning methods. Our
 code is available at \url{https://github.com/apexrl/CoDAIL}.
****

Pseudo-Labeling and Confirmation Bias in Deep Semi-Supervised Learning
Eric Arazo,Diego Ortego,Paul Albert,Noel E. O'Connor,Kevin McGuinness
Semi-supervised learning, i.e. jointly learning from labeled an unlabeled sample
s, is an active research topic due to its key role on relaxing human annotation
constraints. In the context of image classification, recent advances to learn fr
om unlabeled samples are mainly focused on consistency regularization methods th
at encourage invariant predictions for different perturbations of unlabeled sam
ples. We, conversely, propose to learn from unlabeled data by generating soft ps
eudo-labels using the network predictions. We show that a naive pseudo-labeling
overfits to incorrect pseudo-labels due to the so-called confirmation bias and d
emonstrate that mixup augmentation and setting a minimum number of labeled sampl
es per mini-batch are effective regularization techniques for reducing it. The p
roposed approach achieves state-of-the-art results in CIFAR-10/100 and Mini-Imag
eNet despite being much simpler than other state-of-the-art. These results demon
strate that pseudo-labeling can outperform consistency regularization methods, w
hile the opposite was supposed in previous work. Code will be made available.
****

Once-for-All: Train One Network and Specialize it for Efficient Deployment
Han Cai,Chuang Gan,Tianzhe Wang,Zhekai Zhang,Song Han
We address the challenging problem of efficient inference across many devices an
d resource constraints, especially on edge devices. Conventional approaches eit
her manually design or use neural architecture search (NAS) to find a specialize
d neural network and train it from scratch for each case, which is computational
ly prohibitive (causing CO_2 emission as much as 5 cars' lifetime) thus unscal
able. In this work, we propose to train a once-for-all (OFA) network that suppor
ts diverse architectural settings by decoupling training and search, to reduce t
he cost. We can quickly get a specialized sub-network by selecting from the OFA
network without additional training. To efficiently train OFA networks, we also
propose a novel progressive shrinking algorithm, a generalized pruning method th
at reduces the model size across many more dimensions than pruning (depth, width
, kernel size, and resolution). It can obtain a surprisingly large number of sub
-networks ($> 10^{19}$) that can fit different hardware platforms and latency co
nstraints while maintaining the same level of accuracy as training independently

. On diverse edge devices, OFA consistently outperforms state-of-the-art (SOTA) NAS methods (up to 4.0% ImageNet top1 accuracy improvement over MobileNetV3, or same accuracy but 1.5x faster than MobileNetV3, 2.6x faster than EfficientNet w.r.t measured latency) while reducing many orders of magnitude GPU hours and CO_2 emission. In particular, OFA achieves a new SOTA 80.0% ImageNet top-1 accuracy under the mobile setting ($<$600M MACs). OFA is the winning solution for the 3rd Low Power Computer Vision Challenge (LPCVC), DSP classification track and the 4th LPCVC, both classification track and detection track. Code and 50 pre-trained models (for many devices & many latency constraints) are released at https://github.com/mit-han-lab/once-for-all.
****

Generalized Convolutional Forest Networks for Domain Generalization and Visual Recognition
Jongbin Ryu,Gitaek Kwon,Ming-Hsuan Yang,Jongwoo Lim
When constructing random forests, it is of prime importance to ensure high accuracy and low correlation of individual tree classifiers for good performance. Nevertheless, it is typically difficult for existing random forest methods to strike a good balance between these conflicting factors. In this work, we propose a generalized convolutional forest networks to learn a feature space to maximize the strength of individual tree classifiers while minimizing the respective correlation. The feature space is iteratively constructed by a probabilistic triplet sampling method based on the distribution obtained from the splits of the random forest. The sampling process is designed to pull the data of the same label together for higher strength and push away the data frequently falling to the same leaf nodes. We perform extensive experiments on five image classification and two domain generalization datasets with ResNet-50 and DenseNet-161 backbone networks. Experimental results show that the proposed algorithm performs favorably against state-of-the-art methods.
****

Acutum: When Generalization Meets Adaptability
Xunpeng Huang,Zhengyang Liu,Zhe Wang,Yue Yu,Lei Li
In spite of the slow convergence, stochastic gradient descent (SGD) is still the most practical optimization method due to its outstanding generalization ability and simplicity. On the other hand, adaptive methods have attracted much more attention of optimization and machine learning communities, both for the leverage of life-long information and for the deep and fundamental mathematical theory. Taking the best of both worlds is the most exciting and challenging question in the field of optimization for machine learning.

In this paper, we take a small step towards such ultimate goal. We revisit existing adaptive methods from a novel point of view, which reveals a fresh understanding of momentum. Our new intuition empowers us to remove the second moments in Adam without the loss of performance. Based on our view, we propose a new method, named acute adaptive momentum (Acutum). To the best of our knowledge, Acutum is the first adaptive gradient method without second moments. Experimentally, we demonstrate that our method has a faster convergence rate than Adam/Amsgrad, and generalizes as well as SGD with momentum. We also provide a convergence analysis of our proposed method to complement our intuition.
****

FR-GAN: Fair and Robust Training
Yuji Roh,Kangwook Lee,Gyeong Jo Hwang,Steven Euijong Whang,Changho Suh
We consider the problem of fair and robust model training in the presence of data poisoning. Ensuring fairness usually involves a tradeoff against accuracy, so if the data poisoning is mistakenly viewed as additional bias to be fixed, the accuracy will be sacrificed even more. We demonstrate that this phenomenon indeed holds for state-of-the-art model fairness techniques. We then propose FR-GAN, which holistically performs fair and robust model training using generative adversarial networks (GANs). We first use a generator that attempts to classify examples as accurately as possible. In addition, we deploy two discriminators: (1) a fairness discriminator that predicts the sensitive attribute from classification

results and (2) a robustness discriminator that distinguishes examples and predictions from a clean validation set. Our framework respects all the prominent fairness measures: disparate impact, equalized odds, and equal opportunity. Also, FR-GAN optimizes fairness without requiring the knowledge of prior statistics of the sensitive attributes. In our experiments, FR-GAN shows almost no decrease in fairness and accuracy in the presence of data poisoning unlike other state-of-the-art fairness methods, which are vulnerable. In addition, FR-GAN can be adjusted using parameters to maintain reasonable accuracy and fairness even if the validation set is too small or unavailable.
****
SNODE: Spectral Discretization of Neural ODEs for System Identification
Alessio Quaglino,Marco Gallieri,Jonathan Masci,Jan Koutník
This paper proposes the use of spectral element methods \citep{canuto_spectral_1988} for fast and accurate training of Neural Ordinary Differential Equations (ODE-Nets; \citealp{Chen2018NeuralOD}) for system identification. This is achieved by expressing their dynamics as a truncated series of Legendre polynomials. The series coefficients, as well as the network weights, are computed by minimizing the weighted sum of the loss function and the violation of the ODE-Net dynamics. The problem is solved by coordinate descent that alternately minimizes, with respect to the coefficients and the weights, two unconstrained sub-problems using standard backpropagation and gradient methods. The resulting optimization scheme is fully time-parallel and results in a low memory footprint. Experimental comparison to standard methods, such as backpropagation through explicit solvers and the adjoint technique \citep{Chen2018NeuralOD}, on training surrogate models of small and medium-scale dynamical systems shows that it is at least one order of magnitude faster at reaching a comparable value of the loss function. The corresponding testing MSE is one order of magnitude smaller as well, suggesting generalization capabilities increase.
****
Guiding Program Synthesis by Learning to Generate Examples
Larissa Laich,Pavol Bielik,Martin Vechev
A key challenge of existing program synthesizers is ensuring that the synthesized program generalizes well. This can be difficult to achieve as the specification provided by the end user is often limited, containing as few as one or two input-output examples. In this paper we address this challenge via an iterative approach that finds ambiguities in the provided specification and learns to resolve these by generating additional input-output examples. The main insight is to reduce the problem of selecting which program generalizes well to the simpler task of deciding which output is correct. As a result, to train our probabilistic models, we can take advantage of the large amounts of data in the form of program outputs, which are often much easier to obtain than the corresponding ground-truth programs.
****
Fast Neural Network Adaptation via Parameter Remapping and Architecture Search
Jiemin Fang*,Yuzhu Sun*,Kangjian Peng*,Qian Zhang,Yuan Li,Wenyu Liu,Xinggang Wang
Deep neural networks achieve remarkable performance in many computer vision tasks. Most state-of-the-art~(SOTA) semantic segmentation and object detection approaches reuse neural network architectures designed for image classification as the backbone, commonly pre-trained on ImageNet. However, performance gains can be achieved by designing network architectures specifically for detection and segmentation, as shown by recent neural architecture search (NAS) research for detection and segmentation. One major challenge though, is that ImageNet pre-training of the search space representation (a.k.a. super network) or the searched networks incurs huge computational cost. In this paper, we propose a Fast Neural Network Adaptation (FNA) method, which can adapt both the architecture and parameters of a seed network (e.g. a high performing manually designed backbone) to become a network with different depth, width, or kernels via a Parameter Remapping technique, making it possible to utilize NAS for detection/segmentation tasks a lot more efficiently. In our experiments, we conduct FNA on MobileNetV2 to obtain n

ew networks for both segmentation and detection that clearly out-perform existin g networks designed both manually and by NAS. The total computation cost of FNA is significantly less than SOTA segmentation/detection NAS approaches: 1737$\tim es$ less than DPC, 6.8$\times$ less than Auto-DeepLab and 7.4$\times$ less than DetNAS. The code is available at https://github.com/JaminFong/FNA .
****
Measuring Calibration in Deep Learning
Jeremy Nixon,Mike Dusenberry,Ghassen Jerfel,Linchuan Zhang,Dustin Tran
Overconfidence and underconfidence in machine learning classifiers is measured b y calibration: the degree to which the probabilities predicted for each class ma tch the accuracy of the classifier on that prediction. We propose two new measur es for calibration, the Static Calibration Error (SCE) and Adaptive Calibration Error (ACE). These measures take into account every prediction made by a model, in contrast to the popular Expected Calibration Error.
****
R2D2: Reuse & Reduce via Dynamic Weight Diffusion for Training Efficient NLP Mod els
Yi Tay,Aston Zhang,Shuai Zhang,Alvin Chan,Luu Anh Tuan,Siu Cheung Hui
We propose R2D2 layers, a new neural block for training efficient NLP models. Ou r proposed method is characterized by a dynamic weight diffusion mechanism which learns to reuse and reduce parameters in the conventional transformation layer, commonly found in popular Transformer/LSTMs models. Our method is inspired by r ecent Quaternion methods which share parameters via the Hamilton product. This c an be interpreted as a neural and learned approximation of the Hamilton product which imbues our method with increased flexibility and expressiveness, i.e., we are no longer restricted by the 4D nature of Quaternion weight sharing. We condu ct extensive experiments in the NLP domain, showing that R2D2 (i) enables a para meter savings of up to 2 times to 16 times with minimal degradation of performan ce and (ii) outperforms other parameter savings alternative such as low-rank fac torization and Quaternion methods.

****
Exploratory Not Explanatory: Counterfactual Analysis of Saliency Maps for Deep R einforcement Learning
Akanksha Atrey,Kaleigh Clary,David Jensen
Saliency maps are frequently used to support explanations of the behavior of dee p reinforcement learning (RL) agents. However, a review of how saliency maps are used in practice indicates that the derived explanations are often unfalsifiabl e and can be highly subjective. We introduce an empirical approach grounded in c ounterfactual reasoning to test the hypotheses generated from saliency maps and assess the degree to which they correspond to the semantics of RL environments. We use Atari games, a common benchmark for deep RL, to evaluate three types of s aliency maps. Our results show the extent to which existing claims about Atari g ames can be evaluated and suggest that saliency maps are best viewed as an explo ratory tool rather than an explanatory tool.
****
Meta-Learning Deep Energy-Based Memory Models
Sergey Bartunov,Jack Rae,Simon Osindero,Timothy Lillicrap
We study the problem of learning an associative memory model -- a system which i s able to retrieve a remembered pattern based on its distorted or incomplete ver sion.
Attractor networks provide a sound model of associative memory: patterns are sto red as attractors of the network dynamics and associative retrieval is performed by running the dynamics starting from a query pattern until it converges to an attractor.
In such models the dynamics are often implemented as an optimization procedure t hat minimizes an energy function, such as in the classical Hopfield network.
In general it is difficult to derive a writing rule for a given dynamics and ene rgy that is both compressive and fast.
Thus, most research in energy-based memory has been limited either to tractable

energy models not expressive enough to handle complex high-dimensional objects s
uch as natural images, or to models that do not offer fast writing.
We present a novel meta-learning approach to energy-based memory models (EBMM) t
hat allows one to use an arbitrary neural architecture as an energy model and qu
ickly store patterns in its weights.
We demonstrate experimentally that our EBMM approach can build compressed memori
es for synthetic and natural data, and is capable of associative retrieval that
outperforms existing memory systems in terms of the reconstruction error and com
pression rate.
****

Mutual Information Maximization for Robust Plannable Representations
Yiming Ding,Ignasi Clavera,Pieter Abbeel
Extending the capabilities of robotics to real-world complex, unstructured envir
onments requires the capability of developing better perception systems while ma
intaining low sample complexity. When dealing with high-dimensional state spaces
, current methods are either model-free, or model-based with reconstruction base
d objectives. The sample inefficiency of the former constitutes a major barrier
for applying them to the real-world. While the latter present low sample complex
ity, they learn latent spaces that need to reconstruct every single detail of th
e scene. Real-world environments are unstructured and cluttered with objects. Ca
pturing all the variability on the latent representation harms its applicability
 to downstream tasks. In this work, we present mutual information maximization f
or robust plannable representations (MIRO), an information theoretic representat
ional learning objective for model-based reinforcement learning. Our objective o
ptimizes for a latent space that maximizes the mutual information with future ob
servations and emphasizes the relevant aspects of the dynamics, which allows to
capture all the information needed for planning.
We show that our approach learns a latent representation that in cluttered scene
s focuses on the task relevant features, ignoring the irrelevant aspects. At the
 same time, state-of-the-art methods with reconstruction objectives are unable t
o learn in such environments.
****

Depth creates no more spurious local minima in linear networks
Li Zhang
We show that for any convex differentiable loss, a deep linear network has no s
purious local minima as long as it is true for the two layer case. This reducti
on greatly simplifies the study on the existence of spurious local minima in dee
p linear networks. When applied to the quadratic loss, our result immediately im
plies the powerful result by Kawaguchi (2016). Further, with the recent work by
Zhou& Liang (2018), we can remove all the assumptions in (Kawaguchi, 2016). Thi
s property holds for more general “multi-tower” linear networks too. Ou
r proof builds on the work in (Laurent & von Brecht, 2018) and develops a new p
erturbation argument to show that any spurious local minimum must have full rank
, a structural property which can be useful more generally
****

WORD SEQUENCE PREDICTION FOR AMHARIC LANGUAGE
Nuniyat Kifle,Ermias Abebe
Word prediction is guessing what word comes after, based on some current informa
tion, and it is the main
focus of this study. Even though Amharic is used by a large number of population
s, no significant work is
done on the topic. In this study, Amharic word sequence prediction model is deve
loped using Machine
learning. We used statistical methods using Hidden Markov Model by incorporating
 detailed parts of speech
tag and user profiling or adaptation. One of the needs for this research is to o
vercome the challenges on inflected languages. Word sequence prediction is a cha
llenging task for inflected languages (Gustavii &Pettersson, 2003; Seyyed & Assi
, 2005). These kinds of languages are morphologically rich and have enormous wor
d forms, which is a word can

have different forms. As Amharic language is morphologically rich it shares the problem (Tessema,
2014).This problem makes word prediction system much more difficult and results poor performance.
Previous researches used dictionary approach with no consideration of context in formation. Due to this
reason, storing all forms in a dictionary won’t solve the problem as in English and other less inflected
languages. Therefore, we introduced two models; tags and words and linear interp olation that use parts of
speech tag information in addition to word n-grams in order to maximize the like lihood of syntactic
appropriateness of the suggestions. The statistics included in the systems varie s from single word
frequencies to parts-of-speech tag n-grams. We described a combined statistical and lexical word prediction
system and developed Amharic language models of bigram and trigram for the train ing purpose. The overall
study followed Design Science Research Methodology (DSRM).

****

YaoGAN: Learning Worst-case Competitive Algorithms from Self-generated Inputs
Goran Zuzic,Di Wang,Aranyak Mehta,D. Sivakumar
We tackle the challenge of using machine learning to find algorithms with strong worst-case guarantees for online combinatorial optimization problems. Whereas t he previous approach along this direction (Kong et al., 2018) relies on signific ant domain expertise to provide hard distributions over input instances at train ing, we ask whether this can be accomplished from first principles, i.e., withou t any human-provided data beyond specifying the objective of the optimization pr oblem. To answer this question, we draw insights from classic results in game th eory, analysis of algorithms, and online learning to introduce a novel framework . At the high level, similar to a generative adversarial network (GAN), our fram ework has two components whose respective goals are to learn the optimal algorit hm as well as a set of input instances that captures the essential difficulty of the given optimization problem. The two components are trained against each oth er and evolved simultaneously. We test our ideas on the ski rental problem and t he fractional AdWords problem. For these well-studied problems, our preliminary results demonstrate that the framework is capable of finding algorithms as well as difficult input instances that are consistent with known optimal results. We believe our new framework points to a promising direction which can facilitate t he research of algorithm design by leveraging ML to improve the state of the art both in theory and in practice.

****

Annealed Denoising score matching: learning Energy based model in high-dimension al spaces
Zengyi Li,Yubei Chen,Friedrich T. Sommer
Energy based models outputs unmormalized log-probability values given da tasamples. Such a estimation is essential in a variety of application problems suchas sample generation, denoising, sample restoration, outlier detection, Baye sianreasoning, and many more. However, standard maximum likelihood training isc omputationally expensive due to the requirement of sampling model distribution.S core matching potentially alleviates this problem, and denoising score matching(Vincent, 2011) is a particular convenient version. However, previous attemptsf ailed to produce models capable of high quality sample synthesis. We believetha t it is because they only performed denoising score matching over a singlenoise scale. To overcome this limitation, here we instead learn an energy functionusing all noise scales. When sampled using Annealed Langevin dynamics andsingle step denoising jump, our model produced high-quality samples comparabl eto state-of-the-art techniques such as GANs, in addition to assigning likelihoo

d totest data comparable to previous likelihood models. Our model set a new sam
-ple quality baseline in likelihood-based models. We further demonstrate that o
ur model learns sample distribution and generalize well on an image inpainting t
asks.
****
Finding Winning Tickets with Limited (or No) Supervision
Mathilde Caron,Ari Morcos,Piotr Bojanowski,Julien Mairal,Armand Joulin
The lottery ticket hypothesis argues that neural networks contain sparse subnetw
orks, which, if appropriately initialized (the winning tickets), are capable of
matching the accuracy of the full network when trained in isolation. Empirically
 made in different contexts, such an observation opens interesting questions abo
ut the dynamics of neural network optimization and the importance of their initi
alizations. However, the properties of winning tickets are not well understood,
especially the importance of supervision in the generating process. In this pape
r, we aim to answer the following open questions: can we find winning tickets wi
th few data samples or few labels? can we even obtain good tickets without super
vision? Perhaps surprisingly, we provide a positive answer to both, by generatin
g winning tickets with limited access to data, or with self-supervision---thus w
ithout using manual annotations---and then demonstrating the transferability of
the tickets to challenging classification tasks such as ImageNet.

****
Graph Convolutional Reinforcement Learning
Jiechuan Jiang,Chen Dun,Tiejun Huang,Zongqing Lu
Learning to cooperate is crucially important in multi-agent environments. The ke
y is to understand the mutual interplay between agents. However, multi-agent env
ironments are highly dynamic, where agents keep moving and their neighbors chang
e quickly. This makes it hard to learn abstract representations of mutual interp
lay between agents. To tackle these difficulties, we propose graph convolutional
 reinforcement learning, where graph convolution adapts to the dynamics of the u
nderlying graph of the multi-agent environment, and relation kernels capture the
 interplay between agents by their relation representations. Latent features pro
duced by convolutional layers from gradually increased receptive fields are expl
oited to learn cooperation, and cooperation is further improved by temporal rela
tion regularization for consistency. Empirically, we show that our method substa
ntially outperforms existing methods in a variety of cooperative scenarios.
****
Deep Generative Classifier for Out-of-distribution Sample Detection
Dongha Lee,Sehun Yu,Hwanjo Yu
The capability of reliably detecting out-of-distribution samples is one of the k
ey factors in deploying a good classifier, as the test distribution always does
not match with the training distribution in most real-world applications. In thi
s work, we propose a deep generative classifier which is effective to detect out
-of-distribution samples as well as classify in-distribution samples, by integra
ting the concept of Gaussian discriminant analysis into deep neural networks. Un
like the discriminative (or softmax) classifier that only focuses on the decisio
n boundary partitioning its latent space into multiple regions, our generative c
lassifier aims to explicitly model class-conditional distributions as separable
Gaussian distributions. Thereby, we can define the confidence score by the dista
nce between a test sample and the center of each distribution. Our empirical eva
luation on multi-class images and tabular data demonstrate that the generative c
lassifier achieves the best performances in distinguishing out-of-distribution s
amples, and also it can be generalized well for various types of deep neural net
works.
****
Reparameterized Variational Divergence Minimization for Stable Imitation
Dilip Arumugam,Debadeepta Dey,Alekh Agarwal,Asli Celikyilmaz,Elnaz Nouri,Eric Ho
rvitz,Bill Dolan
State-of-the-art results in imitation learning are currently held by adversarial
 methods that iteratively estimate the divergence between student and expert pol

icies and then minimize this divergence to bring the imitation policy closer to expert behavior. Analogous techniques for imitation learning from observations alone (without expert action labels), however, have not enjoyed the same ubiquitous successes.
Recent work in adversarial methods for generative models has shown that the measure used to judge the discrepancy between real and synthetic samples is an algorithmic design choice, and that different choices can result in significant differences in model performance. Choices including Wasserstein distance and various f-divergences have already been explored in the adversarial networks literature, while more recently the latter class has been investigated for imitation learning. Unfortunately, we find that in practice this existing imitation-learning framework for using f-divergences suffers from numerical instabilities stemming from the combination of function approximation and policy-gradient reinforcement learning. In this work, we alleviate these challenges and offer a reparameterization of adversarial imitation learning as f-divergence minimization before further extending the framework to handle the problem of imitation from observations only. Empirically, we demonstrate that our design choices for coupling imitation learning and f-divergences are critical to recovering successful imitation policies. Moreover, we find that with the appropriate choice of f-divergence, we can obtain imitation-from-observation algorithms that outperform baseline approaches and more closely match expert performance in continous-control tasks with low-dimensional observation spaces. With high-dimensional observations, we still observe a significant gap with and without action labels, offering an interesting avenue for future work.
****

Swoosh! Rattle! Thump! - Actions that Sound
Dhiraj Gandhi,Abhinav Gupta,Lerrel Pinto
Truly intelligent agents need to capture the interplay of all their senses to build a rich physical understanding of their world. In robotics, we have seen tremendous progress in using visual and tactile perception; however we have often ignored a key sense: sound. This is primarily due to lack of data that captures the interplay of action and sound. In this work, we perform the first large-scale study of the interactions between sound and robotic action. To do this, we create the largest available sound-action-vision dataset with 15,000 interactions on 60 objects using our robotic platform Tilt-Bot. By tilting objects and allowing them to crash into the walls of a robotic tray, we collect rich four-channel audio information. Using this data, we explore the synergies between sound and action, and present three key insights. First, sound is indicative of fine-grained object class information, e.g., sound can differentiate a metal screwdriver from a metal wrench. Second, sound also contains information about the causal effects of an action, i.e. given the sound produced, we can predict what action was applied on the object. Finally, object representations derived from audio embeddings are indicative of implicit physical properties. We demonstrate that on previously unseen objects, audio embeddings generated through interactions can predict forward models 24% better than passive visual embeddings.
****

Towards Simplicity in Deep Reinforcement Learning: Streamlined Off-Policy Learning
Che Wang,Yanqiu Wu,Quan Vuong,Keith Ross
The field of Deep Reinforcement Learning (DRL) has recently seen a surge in the popularity of maximum entropy reinforcement learning algorithms. Their popularity stems from the intuitive interpretation of the maximum entropy objective and their superior sample efficiency on standard benchmarks. In this paper, we seek to understand the primary contribution of the entropy term to the performance of maximum entropy algorithms. For the Mujoco benchmark, we demonstrate that the entropy term in Soft Actor Critic (SAC) principally addresses the bounded nature of the action spaces. With this insight, we propose a simple normalization scheme which allows a streamlined algorithm without entropy maximization match the performance of SAC. Our experimental results demonstrate a need to revisit the benefits of entropy regularization in DRL. We also propose a simple non-uniform sa

mpling method for selecting transitions from the replay buffer during training. We further show that the streamlined algorithm with the simple non-uniform sampling scheme outperforms SAC and achieves state-of-the-art performance on challenging continuous control tasks.
****

TWIN GRAPH CONVOLUTIONAL NETWORKS: GCN WITH DUAL GRAPH SUPPORT FOR SEMI-SUPERVISED LEARNING

Feng Shi,Yizhou Zhao,Ziheng Xu,Tianyang Liu,Song-Chun Zhu

Graph Neural Networks as a combination of Graph Signal Processing and Deep Convolutional Networks shows great power in pattern recognition in non-Euclidean domains. In this paper, we propose a new method to deploy two pipelines based on the duality of a graph to improve accuracy. By exploring the primal graph and its dual graph where nodes and edges can be treated as one another, we have exploited the benefits of both vertex features and edge features. As a result, we have arrived at a framework that has great potential in both semisupervised and unsupervised learning.
****

Continual Density Ratio Estimation (CDRE): A new method for evaluating generative models in continual learning

Yu Chen,Song Liu,Tom Diethe,Peter Flach

We propose a new method Continual Density Ratio Estimation (CDRE), which can estimate density ratios between a target distribution of real samples and a distribution of samples generated by a model while the model is changing over time and the data of the target distribution is not available after a certain time point. This method perfectly fits the setting of continual learning, in which one model is supposed to learn different tasks sequentially and the most crucial restriction is that model has none or very limited access to the data of all learned tasks. Through CDRE, we can evaluate generative models in continual learning using f-divergences. To the best of our knowledge, there is no existing method that can evaluate generative models under the setting of continual learning without storing real samples from the target distribution.
****

CONTRIBUTION OF INTERNAL REFLECTION IN LANGUAGE EMERGENCE WITH AN UNDER-RESTRICTED SITUATION

Kense Todo,Masayuki Yamamura

Owing to language emergence, human beings have been able to understand the intentions of others, generate common concepts, and extend new concepts. Artificial intelligence researchers have not only predicted words and sentences statistically in machine learning, but also created a language system by communicating with the machine itself. However, strong constraints are exhibited in current studies (supervisor signals and rewards exist, or the concepts were fixed on only a point), thus hindering the emergence of real-world languages. In this study, we improved Batali (1998) and Choi et al. (2018)’s research and attempted language emergence under conditions of low constraints such as human language generation. We included the bias that exists in humans as an “internal reflection function” in to the system. Irrespective of function, messages corresponding to the label could be generated. However, through qualitative and quantitative analysis, we confirmed that the internal reflection function caused “overlearning” and different structuring of message patterns. This result suggested that the internal reflection function performed effectively in creating a grounding language from raw images with an under-restricted situation such as human language generation.
****

Kernelized Wasserstein Natural Gradient

M Arbel,A Gretton,W Li,G Montufar

Many machine learning problems can be expressed as the optimization of some cost functional over a parametric family of probability distributions. It is often beneficial to solve such optimization problems using natural gradient methods. These methods are invariant to the parametrization of the family, and thus can yield more effective optimization. Unfortunately, computing the natural gradient is challenging as it requires inverting a high dimensional matrix at each iteratio

n. We propose a general framework to approximate the natural gradient for the Wa sserstein metric, by leveraging a dual formulation of the metric restricted to a Reproducing Kernel Hilbert Space. Our approach leads to an estimator for gradie nt direction that can trade-off accuracy and computational cost, with theoretica l guarantees. We verify its accuracy on simple examples, and show the advantage of using such an estimator in classification tasks on \texttt{Cifar10} and \text tt{Cifar100} empirically.
****
The Curious Case of Neural Text Degeneration
Ari Holtzman,Jan Buys,Li Du,Maxwell Forbes,Yejin Choi
Despite considerable advances in neural language modeling, it remains an open qu estion what the best decoding strategy is for text generation from a language mo del (e.g. to generate a story). The counter-intuitive empirical observation is t hat even though the use of likelihood as training objective leads to high qualit y models for a broad range of language understanding tasks, maximization-based d ecoding methods such as beam search lead to degeneration — output text that is b land, incoherent, or gets stuck in repetitive loops.

To address this we propose Nucleus Sampling, a simple but effective method to dr aw considerably higher quality text out of neural language models than previous decoding strategies. Our approach avoids text degeneration by truncating the unr eliable tail of the probability distribution, sampling from the dynamic nucleus of tokens containing the vast majority of the probability mass.

To properly examine current maximization-based and stochastic decoding methods, we compare generations from each of these methods to the distribution of human t ext along several axes such as likelihood, diversity, and repetition. Our result s show that (1) maximization is an inappropriate decoding objective for open-end ed text generation, (2) the probability distributions of the best current langua ge models have an unreliable tail which needs to be truncated during generation and (3) Nucleus Sampling is currently the best available decoding strategy for g enerating long-form text that is both high-quality — as measured by human evalua tion — and as diverse as human-written text.
****
Universal approximations of permutation invariant/equivariant functions by deep neural networks
Akiyoshi Sannai,Yuuki Takai,Matthieu Cordonnier
In this paper, we develop a theory about the relationship between G-invariant/ equivariant functions and deep neural networks for finite group G. Especially, for a given G-invariant/equivariant function, we construct its universal appr oximator by deep neural network whose layers equip G-actions and each affine t ransformations are G-equivariant/invariant. Due to representation theory, we c an show that this approximator has exponentially fewer free parameters than usua l models.
****
Improving Robustness Without Sacrificing Accuracy with Patch Gaussian Augmentati on
Raphael Gontijo Lopes,Dong Yin,Ben Poole,Justin Gilmer,Ekin D. Cubuk
Deploying machine learning systems in the real world requires both high accuracy on clean data and robustness to naturally occurring corruptions. While architec tural advances have led to improved accuracy, building robust models remains cha llenging, involving major changes in training procedure and datasets. Prior wor k has argued that there is an inherent trade-off between robustness and accuracy , as exemplified by standard data augmentation techniques such as Cutout, which improves clean accuracy but not robustness, and additive Gaussian noise, which i mproves robustness but hurts accuracy. We introduce Patch Gaussian, a simple aug mentation scheme that adds noise to randomly selected patches in an input image. Models trained with Patch Gaussian achieve state of the art on the CIFAR-10 an d ImageNet Common Corruptions benchmarks while also maintaining accuracy on clea n data. We find that this augmentation leads to reduced sensitivity to high freq

uency noise (similar to Gaussian) while retaining the ability to take advantage of relevant high frequency information in the image (similar to Cutout). We show it can be used in conjunction with other regularization methods and data augmentation policies such as AutoAugment. Finally, we find that the idea of restricting perturbations to patches can also be useful in the context of adversarial learning, yielding models without the loss in accuracy that is found with unconstrained adversarial training.
****

What Can Learned Intrinsic Rewards Capture?
Zeyu Zheng,Junhyuk Oh,Matteo Hessel,Zhongwen Xu,Manuel Kroiss,Hado van Hasselt,David Silver,Satinder Singh
Reinforcement learning agents can include different components, such as policies, value functions, state representations, and environment models. Any or all of these can be the loci of knowledge, i.e., structures where knowledge, whether given or learned, can be deposited and reused. Regardless of its composition, the objective of an agent is behave so as to maximise the sum of suitable scalar functions of state: the rewards. As far as the learning algorithm is concerned, these rewards are typically given and immutable. In this paper we instead consider the proposition that the reward function itself may be a good locus of knowledge. This is consistent with a common use, in the literature, of hand-designed intrinsic rewards to improve the learning dynamics of an agent. We adopt a multi-lifetime setting of the Optimal Rewards Framework, and investigate how meta-learning can be used to find good reward functions in a data-driven way. To this end, we propose to meta-learn an intrinsic reward function that allows agents to maximise their extrinsic rewards accumulated until the end of their lifetimes. This long-term lifetime objective allows our learned intrinsic reward to generate systematic multi-episode exploratory behaviour. Through proof-of-concept experiments, we elucidate interesting forms of knowledge that may be captured by a suitably trained intrinsic reward such as the usefulness of exploring uncertain states and rewards.
****

On Iterative Neural Network Pruning, Reinitialization, and the Similarity of Masks
Michela Paganini,Jessica Forde
We examine how recently documented, fundamental phenomena in deep learn-ing models subject to pruning are affected by changes in the pruning procedure. Specifically, we analyze differences in the connectivity structure and learning dynamics of pruned models found through a set of common iterative pruning techniques, to address questions of uniqueness of trainable, high-sparsity sub-networks, and their dependence on the chosen pruning method. In convolutional layers, we document the emergence of structure induced by magnitude-based un-structured pruning in conjunction with weight rewinding that resembles the effects of structured pruning. We also show empirical evidence that weight stability can be automatically achieved through apposite pruning techniques.
****

Implicit Generative Modeling for Efficient Exploration
Neale Ratzlaff,Qinxun Bai,Li Fuxin,Wei Xu
Efficient exploration remains a challenging problem in reinforcement learning, especially for those tasks where rewards from environments are sparse. A commonly used approach for exploring such environments is to introduce some "intrinsic" reward. In this work, we focus on model uncertainty estimation as an intrinsic reward for efficient exploration. In particular, we introduce an implicit generative modeling approach to estimate a Bayesian uncertainty of the agent's belief of the environment dynamics. Each random draw from our generative model is a neural network that instantiates the dynamic function, hence multiple draws would approximate the posterior, and the variance in the future prediction based on this posterior is used as an intrinsic reward for exploration. We design a training algorithm for our generative model based on the amortized Stein Variational Gradient Descent. In experiments, we compare our implementation with state-of-the-art intrinsic reward-based exploration approaches, including two recent approaches

based on an ensemble of dynamic models. In challenging exploration tasks, our implicit generative model consistently outperforms competing approaches regarding data efficiency in exploration.
****

Continuous Meta-Learning without Tasks
James Harrison,Apoorva Sharma,Chelsea Finn,Marco Pavone
Meta-learning is a promising strategy for learning to efficiently learn within new tasks, using data gathered from a distribution of tasks. However, the meta-learning literature thus far has focused on the task segmented setting, where at train-time, offline data is assumed to be split according to the underlying task, and at test-time, the algorithms are optimized to learn in a single task. In this work, we enable the application of generic meta-learning algorithms to settings where this task segmentation is unavailable, such as continual online learning with a time-varying task. We present meta-learning via online changepoint analysis (MOCA), an approach which augments a meta-learning algorithm with a differentiable Bayesian changepoint detection scheme. The framework allows both training and testing directly on time series data without segmenting it into discrete tasks. We demonstrate the utility of this approach on a nonlinear meta-regression benchmark as well as two meta-image-classification benchmarks.
****

Counterfactual Regularization for Model-Based Reinforcement Learning
Lawrence Neal,Li Fuxin,Xiaoli Fern
In sequential tasks, planning-based agents have a number of advantages over model-free agents, including sample efficiency and interpretability. Recurrent action-conditional latent dynamics models trained from pixel-level observations have been shown to predict future observations conditioned on agent actions accurately enough for planning in some pixel-based control tasks. Typically, models of this type are trained to reconstruct sequences of ground-truth observations, given ground-truth actions. However, an action-conditional model can take input actions and states other than the ground truth, to generate predictions of unobserved counterfactual states. Because counterfactual state predictions are generated by differentiable networks, relationships among counterfactual states can be included in a training objective. We explore the possibilities of counterfactual regularization terms applicable during training of action-conditional sequence models. We evaluate their effect on pixel-level prediction accuracy and model-based agent performance, and we show that counterfactual regularization improves the performance of model-based agents in test-time environments that differ from training.

****

Multilingual Alignment of Contextual Word Representations
Steven Cao,Nikita Kitaev,Dan Klein
We propose procedures for evaluating and strengthening contextual embedding alignment and show that they are useful in analyzing and improving multilingual BERT. In particular, after our proposed alignment procedure, BERT exhibits significantly improved zero-shot performance on XNLI compared to the base model, remarkably matching pseudo-fully-supervised translate-train models for Bulgarian and Greek. Further, to measure the degree of alignment, we introduce a contextual version of word retrieval and show that it correlates well with downstream zero-shot transfer. Using this word retrieval task, we also analyze BERT and find that it exhibits systematic deficiencies, e.g. worse alignment for open-class parts-of-speech and word pairs written in different scripts, that are corrected by the alignment procedure. These results support contextual alignment as a useful concept for understanding large multilingual pre-trained models.
****

A bi-diffusion based layer-wise sampling method for deep learning in large graphs
Yu He,Shiyang Wen,Wenjin Wu,Yan Zhang,Siran Yang,Yuan Wei,Di Zhang,Guojie Song,Wei Lin,Liang Wang,Bo Zheng
The Graph Convolutional Network (GCN) and its variants are powerful models for g

raph representation learning and have recently achieved great success on many gr
aph-based applications. However, most of them target on shallow models (e.g. 2 l
ayers) on relatively small graphs. Very recently, although many acceleration met
hods have been developed for GCNs training,
it still remains a severe challenge how to scale GCN-like models to larger graph
s and deeper layers due to the over-expansion of neighborhoods across layers. In
 this paper, to address the above challenge, we propose a novel layer-wise sampl
ing strategy, which samples the nodes layer by layer conditionally based on the
factors of the bi-directional diffusion between layers. In this way, we potentia
lly restrict the time complexity linear to the number of layers, and construct a
 mini-batch of nodes with high local bi-directional influence (correlation). Fur
ther, we apply the self-attention mechanism to flexibly learn suitable weights f
or the sampled nodes, which allows the model to be able to incorporate both the
first-order and higher-order proximities during a single layer propagation proce
ss without extra recursive propagation or skip connection. Extensive experiments
 on three large benchmark graphs demonstrate the effectiveness and efficiency of
 the proposed model.
****

Learning Video Representations using Contrastive Bidirectional Transformer
Chen Sun,Fabien Baradel,Kevin Murphy,Cordelia Schmid
This paper proposes a self-supervised learning approach for video features that
results in significantly improved performance on downstream tasks (such as video
 classification, captioning and segmentation) compared to existing methods. Our
method extends the BERT model for text sequences to the case of sequences of rea
l-valued feature vectors, by replacing the softmax loss with noise contrastive e
stimation (NCE). We also show how to learn representations from sequences of vis
ual features and sequences of words derived from ASR (automatic speech recogniti
on), and show that such cross-modal training (when possible) helps even more.
****

HUBERT Untangles BERT to Improve Transfer across NLP Tasks
Mehrad Moradshahi,Hamid Palangi,Monica S. Lam,Paul Smolensky,Jianfeng Gao
We introduce HUBERT which combines the structured-representational power of Tens
or-Product Representations (TPRs) and BERT, a pre-trained bidirectional transfor
mer language model. We validate the effectiveness of our model on the GLUE bench
mark and HANS dataset. We also show that there is shared structure between diffe
rent NLP datasets which HUBERT, but not BERT, is able to learn and leverage. Ext
ensive transfer-learning experiments are conducted to confirm this proposition.
****

The Gambler's Problem and Beyond
Baoxiang Wang,Shuai Li,Jiajin Li,Siu On Chan
We analyze the Gambler's problem, a simple reinforcement learning problem where
the gambler has the chance to double or lose their bets until the target is reac
hed. This is an early example introduced in the reinforcement learning textbook
by Sutton and Barto (2018), where they mention an interesting pattern of the opt
imal value function with high-frequency components and repeating non-smooth poin
ts. It is however without further investigation. We provide the exact formula fo
r the optimal value function for both the discrete and the continuous cases. Tho
ugh simple as it might seem, the value function is pathological: fractal, self-s
imilar, derivative taking either zero or infinity, not smooth on any interval, a
nd not written as elementary functions. It is in fact one of the generalized Can
tor functions, where it holds a complexity that has been uncharted thus far. Our
 analyses could lead insights into improving value function approximation, gradi
ent-based algorithms, and Q-learning, in real applications and implementations.
****

GraphAF: a Flow-based Autoregressive Model for Molecular Graph Generation
Chence Shi*,Minkai Xu*,Zhaocheng Zhu,Weinan Zhang,Ming Zhang,Jian Tang
Molecular graph generation is a fundamental problem for drug discovery and has b
een attracting growing attention. The problem is challenging since it requires n
ot only generating chemically valid molecular structures but also optimizing the
ir chemical properties in the meantime. Inspired by the recent progress in deep

generative models, in this paper we propose a flow-based autoregressive model fo
r graph generation called GraphAF. GraphAF combines the advantages of both autor
egressive and flow-based approaches and enjoys: (1) high model flexibility for d
ata density estimation; (2) efficient parallel computation for training; (3) an
iterative sampling process, which allows leveraging chemical domain knowledge fo
r valency checking. Experimental results show that GraphAF is able to generate 6
8\% chemically valid molecules even without chemical knowledge rules and 100\% v
alid molecules with chemical rules. The training process of GraphAF is two times
 faster than the existing state-of-the-art approach GCPN. After fine-tuning the
model for goal-directed property optimization with reinforcement learning, Graph
AF achieves state-of-the-art performance on both chemical property optimization
and constrained property optimization.
****

Off-policy Multi-step Q-learning
Gabriel Kalweit,Maria Huegle,Joschka Boedecker
In the past few years, off-policy reinforcement learning methods have shown prom
ising results in their application for robot control. Deep Q-learning, however,
still suffers from poor data-efficiency which is limiting with regard to real-wo
rld applications. We follow the idea of multi-step TD-learning to enhance data-e
fficiency while remaining off-policy by proposing two novel Temporal-Difference
formulations: (1) Truncated Q-functions which represent the return for the first
 n steps of a policy rollout and (2) Shifted Q-functions, acting as the farsight
ed return after this truncated rollout. We prove that the combination of these s
hort- and long-term predictions is a representation of the full return, leading
to the Composite Q-learning algorithm. We show the efficacy of Composite Q-learn
ing in the tabular case and compare our approach in the function-approximation s
etting with TD3, Model-based Value Expansion and TD3(Delta), which we introduce
as an off-policy variant of TD(Delta). We show on three simulated robot tasks th
at Composite TD3 outperforms TD3 as well as state-of-the-art off-policy multi-st
ep approaches in terms of data-efficiency.
****

Axial Attention in Multidimensional Transformers
Jonathan Ho,Nal Kalchbrenner,Dirk Weissenborn,Tim Salimans
Self-attention effectively captures large receptive fields with high information
 bandwidth, but its computational resource requirements grow quadratically with
the number of points over which attention is performed. For data arranged as lar
ge multidimensional tensors, such as images and videos, the quadratic growth mak
es self-attention prohibitively expensive. These tensors often have thousands of
 positions that one wishes to capture and proposed attentional alternatives eith
er limit the resulting receptive field or require custom subroutines. We propose
 Axial Attention, a simple generalization of self-attention that naturally align
s with the multiple dimensions of the tensors in both the encoding and the decod
ing settings. The Axial Transformer uses axial self-attention layers and a shift
 operation to efficiently build large and full receptive fields. Notably the pr
oposed structure of the layers allows for the vast majority of the context to be
 computed in parallel during decoding without introducing any independence assum
ptions. This semi-parallel structure goes a long way to making decoding from eve
n a very large Axial Transformer broadly applicable. We demonstrate state-of-the
-art results for the Axial Transformer on the ImageNet-32 and ImageNet-64 image
benchmarks as well as on the BAIR Robotic Pushing video benchmark. We open sourc
e the implementation of Axial Transformers.
****

The Surprising Behavior Of Graph Neural Networks
Vivek Kothari,Catherine Tong,Nicholas Lane
We highlight a lack of understanding of the behaviour of Graph Neural Networks (
GNNs) in various topological contexts. We present 4 experimental studies which c
ounter-intuitively demonstrate that the performance of GNNs is weakly dependent
on the topology, sensitive to structural noise and the modality (attributes or e
dges) of information, and degraded by strong coupling between nodal attributes a
nd structure. We draw on the empirical results to recommend reporting of topolog

ical context in GNN evaluation and propose a simple (attribute-structure) decoup
ling method to improve GNN performance.
****
Double Neural Counterfactual Regret Minimization
Hui Li,Kailiang Hu,Shaohua Zhang,Yuan Qi,Le Song
Counterfactual regret minimization (CFR) is a fundamental and effective techniqu
e for solving Imperfect Information Games (IIG). However, the original CFR algor
ithm only works for discrete states and action spaces, and the resulting strateg
y is maintained as a tabular representation. Such tabular representation limits
the method from being directly applied to large games. In this paper, we propose
 a double neural representation for the IIGs, where one neural network represent
s the cumulative regret, and the other represents the average strategy. Such ne
ural representations allow us to avoid manual game abstraction and carry out end
-to-end optimization. To make the learning efficient, we also developed several
novel techniques including a robust sampling method and a mini-batch Monte Carlo
 Counterfactual Regret Minimization (MCCFR) method, which may be of independent
interests. Empirically, on games tractable to tabular approaches, neural strate
gies trained with our algorithm converge comparably to their tabular counterpart
s, and significantly outperform those based on deep reinforcement learning. On
extremely large games with billions of decision nodes, our approach achieved str
ong performance while using hundreds of times less memory than the tabular CFR.
On head-to-head matches of hands-up no-limit texas hold'em, our neural agent bea
t the strong agent ABS-CFR by 9.8 ± 4.1 chips per game. It's a successful appl
ication of neural CFR in large games.

****
Resizable Neural Networks
Yichen Zhu,Xiangyu Zhang,Tong Yang,Jian Sun

In this paper, we present a deep convolutional neural network (CNN) which perfor
ms arbitrary resize operation on intermediate feature map resolution at stage-le
vel. Motivated by weight sharing mechanism in neural architecture search, where
a super-network is trained and sub-networks inherit the weights from the super-n
etwork, we present a novel CNN approach. We construct a spatial super-network wh
ich consists of multiple sub-networks, where each sub-network is a single scale
network that obtain a unique spatial configuration, the convolutional layers are
 shared across all sub-networks. Such network, named as Resizable Neural Network
s, are equivalent to training infinite single scale networks, but has no extra c
omputational cost. Moreover, we present a training algorithm such that all sub-n
etworks achieve better performance than individually trained counterparts. On la
rge-scale ImageNet classification, we demonstrate its effectiveness on various m
odern network architectures such as MobileNet, ShuffleNet, and ResNet.

To go even further, we present three variants of resizable networks: 1) Resizabl
e as Architecture Search (Resizable-NAS). On ImageNet, Resizable-NAS ResNet-50 a
ttain 0.4% higher on accuracy and 44% smaller than the baseline model. 2) Resiza
ble as Data Augmentation (Resizable-Aug). While we use resizable networks as a d
ata augmentation technique, it obtains superior performance on ImageNet classifi
cation, outperform AutoAugment by 1.2% with ResNet-50. 3) Adaptive Resizable Net
work (Resizable-Adapt). We introduce the adaptive resizable networks as dynamic
networks, which further improve the performance with less computational cost via
 data-dependent inference.
****
Multitask Soft Option Learning
Maximilian Igl,Andrew Gambardella,Jinke He,Nantas Nardelli,N. Siddharth,Wendelin
 Böhmer,Shimon Whiteson
We present Multitask Soft Option Learning (MSOL), a hierarchical multi-task fram
ework based on Planning-as-Inference. MSOL extends the concept of Options, using
 separate variational posteriors for each task, regularized by a shared prior. T

he learned soft-options are temporally extended, allowing a higher-level master policy to train faster on new tasks by making decisions with lower frequency. Additionally, MSOL allows fine-tuning of soft-options for new tasks without unlearning previously useful behavior, and avoids problems with local minima in multitask training. We demonstrate empirically that MSOL significantly outperforms both hierarchical and flat transfer-learning baselines in challenging multi-task environments.
****

Adaptive Adversarial Imitation Learning
Yiren Lu,Jonathan Tompson,Sergey Levine
We present the ADaptive Adversarial Imitation Learning (ADAIL) algorithm for learning adaptive policies that can be transferred between environments of varying dynamics, by imitating a small number of demonstrations collected from a single source domain. This problem is important in robotic learning because in real world scenarios 1) reward functions are hard to obtain, 2) learned policies from one domain are difficult to deploy in another due to varying source to target domain statistics, 3) collecting expert demonstrations in multiple environments where the dynamics are known and controlled is often infeasible. We address these constraints by building upon recent advances in adversarial imitation learning; we condition our policy on a learned dynamics embedding and we employ a domain-adversarial loss to learn a dynamics-invariant discriminator. The effectiveness of our method is demonstrated on simulated control tasks with varying environment dynamics and the learned adaptive agent outperforms several recent baselines.
****

Representation Learning with Multisets
Vasco Portilheiro
We study the problem of learning permutation invariant representations that can capture containment relations. We propose training a model on a novel task: predicting the size of the symmetric difference between pairs of multisets, sets which may contain multiple copies of the same object. With motivation from fuzzy set theory, we formulate both multiset representations and how to predict symmetric difference sizes given these representations. We model multiset elements as vectors on the standard simplex and multisets as the summations of such vectors, and we predict symmetric difference as the l1-distance between multiset representations. We demonstrate that our representations more effectively predict the sizes of symmetric differences than DeepSets-based approaches with unconstrained object representations. Furthermore, we demonstrate that the model learns meaningful representations, mapping objects of different classes to different standard basis vectors.
****

Improving Confident-Classifiers For Out-of-distribution Detection
Sachin Vernekar,Ashish Gaurav,Vahdat Abdelzad,Taylor Denouden,Rick Salay,Krzysztof Czarnecki
Discriminatively trained neural classifiers can be trusted, only when the input data comes from the training distribution (in-distribution). Therefore, detecting out-of-distribution (OOD) samples is very important to avoid classification errors. In the context of OOD detection for image classification, one of the recent approaches proposes training a classifier called “confident-classifier” by minimizing the standard cross-entropy loss on in-distribution samples and minimizing the KLdivergence between the predictive distribution of OOD samples in the low-density“boundary” of in-distribution and the uniform distribution (maximizing the entropy of the outputs). Thus, the samples could be detected as OOD if they have low confidence or high entropy. In this paper, we analyze this setting both theoretically and experimentally. We also propose a novel algorithm to generate the“boundary” OOD samples to train a classifier with an explicit “reject” class for OOD samples. We compare our approach against several recent classifier-based OOD detectors including the confident-classifiers on MNIST and Fashion-MNISTdatasets. Overall the proposed approach consistently performs better than others across most of the experiments.
****

Cyclic Graph Dynamic Multilayer Perceptron for Periodic Signals
Mikio Furokawa,Erik Gest,Takayuki Hirano,Kamal Youcef-Toumi
We propose a feature extraction for periodic signals. Virtually every mechanized transportation vehicle, power generation, industrial machine, and robotic system contains rotating shafts. It is possible to collect data about periodicity by mea- suring a shaft’s rotation. However, it is difficult to perfectly control the collection timing of the measurements. Imprecise timing creates phase shifts in the resulting data. Although a phase shift does not materially affect the measurement of any given data point collected, it does alter the order in which all of the points are col- lected. It is difficult for classical methods, like multi-layer perceptron, to identify or quantify these alterations because they depend on the order of the input vectors’ components. This paper proposes a robust method for extracting features from phase shift data by adding a graph structure to each data point and constructing a suitable machine learning architecture for graph data with cyclic permutation. Simulation and experimental results illustrate its effectiveness.
****
Accelerating Monte Carlo Bayesian Inference via Approximating Predictive Uncertainty over the Simplex
Yufei Cui,Wuguannan Yao,Qiao Li,Antoni Chan,Chun Jason Xue
Estimating the predictive uncertainty of a Bayesian learning model is critical in various decision-making problems, e.g., reinforcement learning, detecting adversarial attack, self-driving car. As the model posterior is almost always intractable, most efforts were made on finding an accurate approximation the true posterior. Even though a decent estimation of the model posterior is obtained, another approximation is required to compute the predictive distribution over the desired output. A common accurate solution is to use Monte Carlo (MC) integration. However, it needs to maintain a large number of samples, evaluate the model repeatedly and average multiple model outputs. In many real-world cases, this is computationally prohibitive. In this work, assuming that the exact posterior or a decent approximation is obtained, we propose a generic framework to approximate the output probability distribution induced by model posterior with a parameterized model and in an amortized fashion. The aim is to approximate the true uncertainty of a specific Bayesian model, meanwhile alleviating the heavy workload of MC integration at testing time. The proposed method is universally applicable to Bayesian classification models that allow for posterior sampling. Theoretically, we show that the idea of amortization incurs no additional costs on approximation performance. Empirical results validate the strong practical performance of our approach.
****
Certifiably Robust Interpretation in Deep Learning
Alexander Levine,Sahil Singla,Soheil Feizi
Deep learning interpretation is essential to explain the reasoning behind model predictions. Understanding the robustness of interpretation methods is important especially in sensitive domains such as medical applications since interpretation results are often used in downstream tasks. Although gradient-based saliency maps are popular methods for deep learning interpretation, recent works show that they can be vulnerable to adversarial attacks. In this paper, we address this problem and provide a certifiable defense method for deep learning interpretation. We show that a sparsified version of the popular SmoothGrad method, which computes the average saliency maps over random perturbations of the input, is certifiably robust against adversarial perturbations. We obtain this result by extending recent bounds for certifiably robust smooth classifiers to the interpretation setting. Experiments on ImageNet samples validate our theory.
****
Continuous Convolutional Neural Network forNonuniform Time Series
Hui Shi,Yang Zhang,Hao Wu,Shiyu Chang,Kaizhi Qian,Mark Hasegawa-Johnson,Jishen Zhao
Convolutional neural network (CNN) for time series data implicitly assumes that the data are uniformly sampled, whereas many event-based and multi-modal data ar

e nonuniform or have heterogeneous sampling rates. Directly applying regularCNN to nonuniform time series is ungrounded, because it is unable to recognize and extract common patterns from the nonuniform input signals. Converting the nonuniform time series to uniform ones by interpolation preserves the pattern extraction capability of CNN, but the interpolation kernels are often preset and may be unsuitable for the data or tasks. In this paper, we propose the ContinuousCNN (CCNN), which estimates the inherent continuous inputs by interpolation, and performs continuous convolution on the continuous input. The interpolation and convolution kernels are learned in an end-to-end manner, and are able to learn useful patterns despite the nonuniform sampling rate. Besides, CCNN is a strict generalization to CNN. Results of several experiments verify that CCNN achieves abetter performance on nonuniform data, and learns meaningful continuous kernels
****

DS-VIC: Unsupervised Discovery of Decision States for Transfer in RL
Nirbhay Modhe,Prithvijit Chattopadhyay,Mohit Sharma,Abhishek Das,Devi Parikh,Dhruv Batra,Ramakrishna Vedantam
We learn to identify decision states, namely the parsimonious set of states where decisions meaningfully affect the future states an agent can reach in an environment. We utilize the VIC framework, which maximizes an agent's `empowerment', ie the ability to reliably reach a diverse set of states -- and formulate a sandwich bound on the empowerment objective that allows identification of decision states. Unlike previous work, our decision states are discovered without extrinsic rewards -- simply by interacting with the world. Our results show that our decision states are: 1) often interpretable, and 2) lead to better exploration on downstream goal-driven tasks in partially observable environments.
****

Neural Policy Gradient Methods: Global Optimality and Rates of Convergence
Lingxiao Wang,Qi Cai,Zhuoran Yang,Zhaoran Wang
Policy gradient methods with actor-critic schemes demonstrate tremendous empirical successes, especially when the actors and critics are parameterized by neural networks. However, it remains less clear whether such "neural" policy gradient methods converge to globally optimal policies and whether they even converge at all. We answer both the questions affirmatively in the overparameterized regime. In detail, we prove that neural natural policy gradient converges to a globally optimal policy at a sublinear rate. Also, we show that neural vanilla policy gradient converges sublinearly to a stationary point. Meanwhile, by relating the suboptimality of the stationary points to the~representation power of neural actor and critic classes, we prove the global optimality of all stationary points under mild regularity conditions. Particularly, we show that a key to the global optimality and convergence is the "compatibility" between the actor and critic, which is ensured by sharing neural architectures and random initializations across the actor and critic. To the best of our knowledge, our analysis establishes the first global optimality and convergence guarantees for neural policy gradient methods.
****

Multi-objective Neural Architecture Search via Predictive Network Performance Optimization
Han Shi,Renjie Pi,Hang Xu,Zhenguo Li,James T. Kwok,Tong Zhang
Neural Architecture Search (NAS) has shown great potentials in finding a better neural network design than human design. Sample-based NAS is the most fundamental method aiming at exploring the search space and evaluating the most promising architecture. However, few works have focused on improving the sampling efficiency for a multi-objective NAS. Inspired by the nature of the graph structure of a neural network, we propose BOGCN-NAS, a NAS algorithm using Bayesian Optimization with Graph Convolutional Network (GCN) predictor. Specifically, we apply GCN as a surrogate model to adaptively discover and incorporate nodes structure to approximate the performance of the architecture. For NAS-oriented tasks, we also design a weighted loss focusing on architectures with high performance. Our method further considers an efficient multi-objective search which can be flexibly injected into any sample-based NAS pipelines to efficiently find the best speed/a

ccuracy trade-off. Extensive experiments are conducted to verify the effectiveness of our method over many competing methods, e.g. 128.4x more efficient than Random Search and 7.8x more efficient than previous SOTA LaNAS for finding the best architecture on the largest NAS dataset NasBench-101.
****
Triple Wins: Boosting Accuracy, Robustness and Efficiency Together by Enabling Input-Adaptive Inference
Ting-Kuei Hu,Tianlong Chen,Haotao Wang,Zhangyang Wang
Deep networks were recently suggested to face the odds between accuracy (on clean natural images) and robustness (on adversarially perturbed images) (Tsipras et al., 2019). Such a dilemma is shown to be rooted in the inherently higher sample complexity (Schmidt et al., 2018) and/or model capacity (Nakkiran, 2019), for learning a high-accuracy and robust classifier. In view of that, give a classification task, growing the model capacity appears to help draw a win-win between accuracy and robustness, yet at the expense of model size and latency, therefore posing challenges for resource-constrained applications. Is it possible to co-design model accuracy, robustness and efficiency to achieve their triple wins? This paper studies multi-exit networks associated with input-adaptive efficient inference, showing their strong promise in achieving a “sweet point" in co-optimizing model accuracy, robustness, and efficiency. Our proposed solution, dubbed Robust Dynamic Inference Networks (RDI-Nets), allows for each input (either clean or adversarial) to adaptively choose one of the multiple output layers (early branches or the final one) to output its prediction. That multi-loss adaptivity adds new variations and flexibility to adversarial attacks and defenses, on which we present a systematical investigation. We show experimentally that by equipping existing backbones with such robust adaptive inference, the resulting RDI-Nets can achieve better accuracy and robustness, yet with over 30% computational savings, compared to the defended original models.

****
Dynamic Sparse Training: Find Efficient Sparse Network From Scratch With Trainable Masked Layers
Junjie LIU,Zhe XU,Runbin SHI,Ray C. C. Cheung,Hayden K.H. So
We present a novel network pruning algorithm called Dynamic Sparse Training that can jointly ■nd the optimal network parameters and sparse network structure in a uni■ed optimization process with trainable pruning thresholds. These thresholds can have ■ne-grained layer-wise adjustments dynamically via backpropagation. We demonstrate that our dynamic sparse training algorithm can easily train very sparse neural network models with little performance loss using the same training epochs as dense models. Dynamic Sparse Training achieves prior art performance compared with other sparse training algorithms on various network architectures. Additionally, we have several surprising observations that provide strong evidence to the effectiveness and ef■ciency of our algorithm. These observations reveal the underlying problems of traditional three-stage pruning algorithms and present the potential guidance provided by our algorithm to the design of more compact network architectures.
****
A Mean-Field Theory for Kernel Alignment with Random Features in Generative Adverserial Networks
Masoud Badiei Khuzani,Liyue Shen,Shahin Shahrampour,Lei Xing
We propose a novel supervised learning method to optimize the kernel in maximum mean discrepancy generative adversarial networks (MMD GANs). Specifically, we characterize a distributionally robust optimization problem to compute a good distribution for the random feature model of Rahimi and Recht to approximate a good kernel function. Due to the fact that the distributional optimization is infinite dimensional, we consider a Monte-Carlo sample average approximation (SAA) to obtain a more tractable finite dimensional optimization problem. We subsequently leverage a particle stochastic gradient descent (SGD) method to solve finite dimensional optimization problems. Based on a mean-field analysis, we then prove that the empirical distribution of the interactive particles system at each iterat

ion of the SGD follows the path of the gradient descent flow on the Wasserstein manifold. We also establish the non-asymptotic consistency of the finite sample estimator. Our empirical evaluation on synthetic data-set as well as MNIST and CIFAR-10 benchmark data-sets indicates that our proposed MMD GAN model with kernel learning indeed attains higher inception scores well as Fr\`{e}chet inception distances and generates better images compared to the generative moment matching network (GMMN) and MMD GAN with untrained kernels.
****
Learning Key Steps to Attack Deep Reinforcement Learning Agents
Chien-Min Yu,Hsuan-Tien Lin
Deep reinforcement learning agents are known to be vulnerable to adversarial attacks. In particular, recent studies have shown that attacking a few key steps is effective for decreasing the agent's cumulative reward. However, all existing attacking methods find those key steps with human-designed heuristics, and it is not clear how more effective key steps can be identified. This paper introduces a novel reinforcement learning framework that learns more effective key steps through interacting with the agent. The proposed framework does not require any human heuristics nor knowledge, and can be flexibly coupled with any white-box or black-box adversarial attack scenarios. Experiments on benchmark Atari games across different scenarios demonstrate that the proposed framework is superior to existing methods for identifying more effective key steps.
****
Beyond Linearization: On Quadratic and Higher-Order Approximation of Wide Neural Networks
Yu Bai,Jason D. Lee
Recent theoretical work has established connections between over-parametrized neural networks and linearized models governed by the Neural Tangent Kernels (NTKs). NTK theory leads to concrete convergence and generalization results, yet the empirical performance of neural networks are observed to exceed their linearized models, suggesting insufficiency of this theory.
Towards closing this gap, we investigate the training of over-parametrized neural networks that are beyond the NTK regime yet still governed by the Taylor expansion of the network. We bring forward the idea of randomizing the neural networks, which allows them to escape their NTK and couple with quadratic models. We show that the optimization landscape of randomized two-layer networks are nice and amenable to escaping-saddle algorithms. We prove concrete generalization and expressivity results on these randomized networks, which lead to sample complexity bounds (of learning certain simple functions) that match the NTK and can in addition be better by a dimension factor when mild distributional assumptions are present. We demonstrate that our randomization technique can be generalized systematically beyond the quadratic case, by using it to find networks that are coupled with higher-order terms in their Taylor series.

****
On PAC-Bayes Bounds for Deep Neural Networks using the Loss Curvature
Konstantinos Pitas
We investigate whether it's possible to tighten PAC-Bayes bounds for deep neural networks by utilizing the Hessian of the training loss at the minimum. For the case of Gaussian priors and posteriors we introduce a Hessian-based method to obtain tighter PAC-Bayes bounds that relies on closed form solutions of layerwise subproblems. We thus avoid commonly used variational inference techniques which can be difficult to implement and time consuming for modern deep architectures. We conduct a theoretical analysis that links the random initialization, minimum, and curvature at the minimum of a deep neural network to limits on what is provable about generalization through PAC-Bayes. Through careful experiments we validate our theoretical predictions and analyze the influence of the prior mean, prior covariance, posterior mean and posterior covariance on obtaining tighter bounds.
****
Deep Graph Matching Consensus

Matthias Fey,Jan E. Lenssen,Christopher Morris,Jonathan Masci,Nils M. Kriege
This work presents a two-stage neural architecture for learning and refining structural correspondences between graphs. First, we use localized node embeddings computed by a graph neural network to obtain an initial ranking of soft correspondences between nodes. Secondly, we employ synchronous message passing networks to iteratively re-rank the soft correspondences to reach a matching consensus in local neighborhoods between graphs. We show, theoretically and empirically, that our message passing scheme computes a well-founded measure of consensus for corresponding neighborhoods, which is then used to guide the iterative re-ranking process. Our purely local and sparsity-aware architecture scales well to large, real-world inputs while still being able to recover global correspondences consistently. We demonstrate the practical effectiveness of our method on real-world tasks from the fields of computer vision and entity alignment between knowledge graphs, on which we improve upon the current state-of-the-art.
****

Self-Supervised Learning of Appliance Usage
Chen-Yu Hsu,Abbas Zeitoun,Guang-He Lee,Dina Katabi,Tommi Jaakkola
Learning home appliance usage is important for understanding people's activities and optimizing energy consumption. The problem is modeled as an event detection task, where the objective is to learn when a user turns an appliance on, and which appliance it is (microwave, hair dryer, etc.). Ideally, we would like to solve the problem in an unsupervised way so that the method can be applied to new homes and new appliances without any labels. To this end, we introduce a new deep learning model that takes input from two home sensors: 1) a smart electricity meter that outputs the total energy consumed by the home as a function of time, and 2) a motion sensor that outputs the locations of the residents over time. The model learns the distribution of the residents' locations conditioned on the home energy signal. We show that this cross-modal prediction task allows us to detect when a particular appliance is used, and the location of the appliance in the home, all in a self-supervised manner, without any labeled data.
****

Gaussian Conditional Random Fields for Classification
Andrija Petrovic,Mladen Nikolic,Milos Jovanovic,Boris Delibasic
In this paper, a Gaussian conditional random field model for structured binary classification (GCRFBC) is proposed. The model is applicable to classification problems with undirected graphs, intractable for standard classification CRFs. The model representation of GCRFBC is extended by latent variables which yield some appealing properties. Thanks to the GCRF latent structure, the model becomes tractable, efficient, and open to improvements previously applied to GCRF regression. Two different forms of the algorithm are presented: GCRFBCb (GCRGBC - Bayesian) and GCRFBCnb (GCRFBC - non-Bayesian). The extended method of local variational approximation of sigmoid function is used for solving empirical Bayes in GCRFBCb variant, whereas MAP value of latent variables is the basis for learning and inference in the GCRFBCnb variant. The inference in GCRFBCb is solved by Newton-Cotes formulas for one-dimensional integration. Both models are evaluated on synthetic data and real-world data. It was shown that both models achieve better prediction performance than relevant baselines. Advantages and disadvantages of the proposed models are discussed.
****

Fourier networks for uncertainty estimates and out-of-distribution detection
Hartmut Maennel,Alexandru ■ifrea
A simple method for obtaining uncertainty estimates for Neural Network classifiers (e.g. for out-of-distribution detection) is to use an ensemble of independently trained networks and average the softmax outputs. While this method works, its results are still very far from human performance on standard data sets. We investigate how this method works and observe three fundamental limitations: "Unreasonable" extrapolation, "unreasonable" agreement between the networks in an ensemble, and the filtering out of features that distinguish the training distribution from some out-of-distribution inputs, but do not contribute to the classification. To mitigate these problems we suggest "large" initializations in the firs

t layers and changing the activation function to sin(x) in the last hidden layer
. We show that this combines the out-of-distribution behavior from nearest neigh
bor methods with the generalization capabilities of neural networks, and achieve
s greatly improved out-of-
distribution detection on standard data sets (MNIST/fashionMNIST/notMNIST, SVHN/
CIFAR10).
****

Semantic Hierarchy Emerges in the Deep Generative Representations for Scene Synt
hesis
Ceyuan Yang,Yujun Shen,Bolei Zhou
Despite the success of Generative Adversarial Networks (GANs) in image synthesis
, there lacks enough understanding on what networks have learned inside the deep
 generative representations and how photo-realistic images are able to be compos
ed from random noises. In this work, we show that highly-structured semantic hie
rarchy emerges from the generative representations as the variation factors for
synthesizing scenes. By probing the layer-wise representations with a broad set
of visual concepts at different abstraction levels, we are able to quantify the
causality between the activations and the semantics occurring in the output imag
e. Such a quantification identifies the human-understandable variation factors l
earned by GANs to compose scenes. The qualitative and quantitative results sugge
st that the generative representations learned by GAN are specialized to synthes
ize different hierarchical semantics: the early layers tend to determine the spa
tial layout and configuration, the middle layers control the categorical objects
, and the later layers finally render the scene attributes as well as color sche
me. Identifying such a set of manipulatable latent semantics facilitates semanti
c scene manipulation.
****

Quantum Algorithms for Deep Convolutional Neural Networks
Iordanis Kerenidis,Jonas Landman,Anupam Prakash
Quantum computing is a powerful computational paradigm with applications in seve
ral fields, including machine learning. In the last decade, deep learning, and i
n particular Convolutional Neural Networks (CNN), have become essential for appl
ications in signal processing and image recognition. Quantum deep learning, howe
ver, remains a challenging problem, as it is difficult to implement non linearit
ies with quantum unitaries. In this paper we propose a quantum algorithm for eva
luating and training deep convolutional neural networks with potential speedups
over classical CNNs for both the forward and backward passes. The quantum CNN (Q
CNN) reproduces completely the outputs of the classical CNN and allows for non l
inearities and pooling operations. The QCNN is in particular interesting for dee
p networks and could allow new frontiers in the image recognition domain, by all
owing for many more convolution kernels, larger kernels, high dimensional inputs
 and high depth input channels. We also present numerical simulations for the cl
assification of the MNIST dataset to provide practical evidence for the efficien
cy of the QCNN.
****

TWO-STEP UNCERTAINTY NETWORK FOR TASKDRIVEN SENSOR PLACEMENT
Yangyang Sun,Yang Zhang,Hassan Foroosh,Shuo Pang
Optimal sensor placement achieves the minimal cost of sensors while obtaining th
e prespecified objectives. In this work, we propose a framework for sensor place
ment to maximize the information gain called Two-step Uncertainty Network(TUN).
TUN encodes an arbitrary number of measurements, models the conditional distribu
tion of high dimensional data, and estimates the task-specific information gain
at un-observed locations. Experiments on the synthetic data show that TUN outper
forms the random sampling strategy and Gaussian Process-based strategy consisten
tly.
****

EXPLOITING SEMANTIC COHERENCE TO IMPROVE PREDICTION IN SATELLITE SCENE IMAGE ANA
LYSIS: APPLICATION TO DISEASE DENSITY ESTIMATION
Rahman Sanya,Gilbert Maiga,Ernest Mwebaze
High intra-class diversity and inter-class similarity is a characteristic of rem

ote sensing scene image data sets currently posing significant difficulty for de
ep learning algorithms on classification tasks. To improve accuracy, post-classi
fication
methods have been proposed for smoothing results of model predictions. However,
those approaches require an additional neural network to perform the smoothing o
peration, which adds overhead to the task. We propose an approach that involves
learning deep features directly over neighboring scene images without requiring
use of a cleanup model. Our approach utilizes a siamese network to improve the d
iscriminative power of convolutional neural networks on a pair
of neighboring scene images. It then exploits semantic coherence between this pa
ir to enrich the feature vector of the image for which we want to predict a labe
l.
Empirical results show that this approach provides a viable alternative to exist
ing methods. For example, our model improved prediction accuracy by 1 percentage
 point and dropped the mean squared error value by 0.02 over the baseline, on a
disease density estimation task. These performance gains are comparable with res
ults from existing post-classification methods, moreover without implementation
overheads.
****

Deep Batch Active Learning by Diverse, Uncertain Gradient Lower Bounds
Jordan T. Ash,Chicheng Zhang,Akshay Krishnamurthy,John Langford,Alekh Agarwal
We design a new algorithm for batch active learning with deep neural network mod
els. Our algorithm, Batch Active learning by Diverse Gradient Embeddings (BADGE)
, samples groups of points that are disparate and high-magnitude when represente
d in a hallucinated gradient space, a strategy designed to incorporate both pred
ictive uncertainty and sample diversity into every selected batch. Crucially, BA
DGE trades off between diversity and uncertainty without requiring any hand-tune
d hyperparameters. While other approaches sometimes succeed for particular batch
 sizes or architectures, BADGE consistently performs as well or better, making i
t a useful option for real world active learning problems.
****

Abstractive Dialog Summarization with Semantic Scaffolds
Lin Yuan,Zhou Yu
The demand for abstractive dialog summary is growing in real-world applications.
 For example, customer service center or hospitals would like to summarize custo
mer service interaction and doctor-patient interaction. However, few researchers
 explored abstractive summarization on dialogs due to the lack of suitable data
sets. We propose an abstractive dialog summarization dataset based on MultiWOZ.
If we directly apply previous state-of-the-art document summarization methods on
 dialogs, there are two significant drawbacks: the informative entities such as
restaurant names are difficult to preserve, and the contents from different dial
og domains are sometimes mismatched. To address these two drawbacks, we propose
Scaffold Pointer Network (SPNet) to utilize the existing annotation on speaker r
ole, semantic slot and dialog domain. SPNet incorporates these semantic scaffold
s for dialog summarization. Since ROUGE cannot capture the two drawbacks mention
ed, we also propose a new evaluation metric that considers critical informative
entities in the text. On MultiWOZ, our proposed SPNet outperforms state-of-the-a
rt abstractive summarization methods on all the automatic and human evaluation m
etrics.
****

Evaluating Semantic Representations of Source Code
Yaza Wainakh,Moiz Rauf,Michael Pradel
Learned representations of source code enable various software developer tools,
e.g., to detect bugs or to predict program properties. At the core of code repre
sentations often are word embeddings of identifier names in source code, because
 identifiers account for the majority of source code vocabulary and convey impor
tant semantic information. Unfortunately, there currently is no generally accept
ed way of evaluating the quality of word embeddings of identifiers, and current
evaluations are biased toward specific downstream tasks. This paper presents IdB
ench, the first benchmark for evaluating to what extent word embeddings of ident

ifiers represent semantic relatedness and similarity. The benchmark is based on thousands of ratings gathered by surveying 500 software developers. We use IdBench to evaluate state-of-the-art embedding techniques proposed for natural language, an embedding technique specifically designed for source code, and lexical string distance functions, as these are often used in current developer tools. Our results show that the effectiveness of embeddings varies significantly across different embedding techniques and that the best available embeddings successfully represent semantic relatedness. On the downside, no existing embedding provides a satisfactory representation of semantic similarities, e.g., because embeddings consider identifiers with opposing meanings as similar, which may lead to fatal mistakes in downstream developer tools. IdBench provides a gold standard to guide the development of novel embeddings that address the current limitations.

****
Searching to Exploit Memorization Effect in Learning from Corrupted Labels
Hansi Yang,Quanming Yao,Bo Han,Gang Niu
Sample-selection approaches, which attempt to pick up clean instances from the training data set, have become one promising direction to robust learning from corrupted labels. These methods all build on the memorization effect, which means deep networks learn easy patterns first and then gradually over-fit the training data set. In this paper, we show how to properly select instances so that the training process can benefit the most from the memorization effect is a hard problem. Specifically, memorization can heavily depend on many factors, e.g., data set and network architecture. Nonetheless, there still exists general patterns of how memorization can occur. These facts motivate us to exploit memorization by automated machine learning (AutoML) techniques. First, we designed an expressive but compact search space based on observed general patterns. Then, we propose to use the natural gradient-based search algorithm to efficiently search through space. Finally, extensive experiments on both synthetic data sets and benchmark data sets demonstrate that the proposed method can not only be much efficient than existing AutoML algorithms but can also achieve much better performance than the state-of-the-art approaches for learning from corrupted labels.
****
Understanding l4-based Dictionary Learning: Interpretation, Stability, and Robustness
Yuexiang Zhai,Hermish Mehta,Zhengyuan Zhou,Yi Ma
Recently, the ℓ^4-norm maximization has been proposed to solve the sparse dictionary learning (SDL) problem. The simple MSP (matching, stretching, and projection) algorithm proposed by \cite{zhai2019a} has proved surprisingly efficient and effective. This paper aims to better understand this algorithm from its strong geometric and statistical connections with the classic PCA and ICA, as well as their associated fixed-point style algorithms. Such connections provide a unified way of viewing problems that pursue {\em principal}, {\em independent}, or {\em sparse} components of high-dimensional data. Our studies reveal additional good properties of ℓ^4-maximization: not only is the MSP algorithm for sparse coding insensitive to small noise, but it is also robust to outliers and resilient to sparse corruptions. We provide statistical justification for such inherently nice properties. To corroborate the theoretical analysis, we also provide extensive and compelling experimental evidence with both synthetic data and real images.
****
Balancing Cost and Benefit with Tied-Multi Transformers
Raj Dabre,Raphael Rubino,Atsushi Fujita
This paper proposes a novel procedure for training multiple Transformers with tied parameters which compresses multiple models into one enabling the dynamic choice of the number of encoder and decoder layers during decoding. In sequence-to-sequence modeling, typically, the output of the last layer of the N-layer encoder is fed to the M-layer decoder, and the output of the last decoder layer is used to compute loss. Instead, our method computes a single loss consisting of NxM losses, where each loss is computed from the output of one of the M decoder laye

rs connected to one of the N encoder layers. A single model trained by our metho d subsumes multiple models with different number of encoder and decoder layers, and can be used for decoding with fewer than the maximum number of encoder and d ecoder layers. We then propose a mechanism to choose a priori the number of enco der and decoder layers for faster decoding, and also explore recurrent stacking of layers and knowledge distillation to enable further parameter reduction. In a case study of neural machine translation, we present a cost-benefit analysis of the proposed approaches and empirically show that they greatly reduce decoding costs while preserving translation quality.
****

BOSH: An Efficient Meta Algorithm for Decision-based Attacks
Zhenxin Xiao,Puyudi Yang,Yuchen Jiang,Kai-Wei Chang,Cho-Jui Hsieh
Adversarial example generation becomes a viable method for evaluating the robust ness of a machine learning model. In this paper, we consider hard-label black- b ox attacks (a.k.a. decision-based attacks), which is a challenging setting that generates adversarial examples based on only a series of black-box hard-label qu eries. This type of attacks can be used to attack discrete and complex models, s uch as Gradient Boosting Decision Tree (GBDT) and detection-based defense models . Existing decision-based attacks based on iterative local updates often get stu ck in a local minimum and fail to generate the optimal adversarial example with the smallest distortion. To remedy this issue, we propose an efficient meta algo rithm called BOSH-attack, which tremendously improves existing algorithms throug h Bayesian Optimization (BO) and Successive Halving (SH). In particular, instead of traversing a single solution path when searching an adversarial example, we maintain a pool of solution paths to explore important regions. We show empirica lly that the proposed algorithm converges to a better solution than existing app roaches, while the query count is smaller than applying multiple random initiali zations by a factor of 10.
****

MGP-AttTCN: An Interpretable Machine Learning Model for the Prediction of Sepsis
Margherita Rosnati,Vincent Fortuin
With a mortality rate of 5.4 million lives worldwide every year and a healthcare cost of more than 16 billion dollars in the USA alone, sepsis is one of the lea ding causes of hospital mortality and an increasing concern in the ageing wester n world. Recently, medical and technological advances have helped re-define the illness criteria of this disease, which is otherwise poorly understood by the me dical society. Together with the rise of widely accessible Electronic Health Rec ords, the advances in data mining and complex nonlinear algorithms are a promisi ng avenue for the early detection of sepsis. This work contributes to the resear ch effort in the field of automated sepsis detection with an open-access labelli ng of the medical MIMIC-III data set. Moreover, we propose MGP-AttTCN: a joint m ultitask Gaussian Process and attention-based deep learning model to early predi ct the occurrence of sepsis in an interpretable manner. We show that our model o utperforms the current state-of-the-art and present evidence that different labe lling heuristics lead to discrepancies in task difficulty.
****

Unsupervised Representation Learning by Predicting Random Distances
Hu Wang,Guansong Pang,Chunhua Shen,Congbo Ma
Deep neural networks have gained tremendous success in a broad range of machine learning tasks due to its remarkable capability to learn semantic-rich features from high-dimensional data. However, they often require large-scale labelled dat a to successfully learn such features, which significantly hinders their adaptio n into unsupervised learning tasks, such as anomaly detection and clustering, an d limits their applications into critical domains where obtaining massive labell ed data is prohibitively expensive. To enable downstream unsupervised learning o n those domains, in this work we propose to learn features without using any lab elled data by training neural networks to predict data distances in a randomly p rojected space. Random mapping is a highly efficient yet theoretical proven appr oach to obtain approximately preserved distances. To well predict these random d istances, the representation learner is optimised to learn class structures that

are implicitly embedded in the randomly projected space. Experimental results o
n 19 real-world datasets show our learned representations substantially outperfo
rm state-of-the-art competing methods in both anomaly detection and clustering t
asks.
****

ConQUR: Mitigating Delusional Bias in Deep Q-Learning

DiJia-Andy Su,Jayden Ooi,Tyler Lu,Dale Schuurmans,Craig Boutilier■

Delusional bias is a fundamental source of error in approximate Q-learning. To d
ate, the only techniques that explicitly address delusion require comprehensive
search using tabular value estimates. In this paper, we develop efficient method
s to mitigate delusional bias by training Q-approximators with labels that are "
consistent" with the underlying greedy policy class. We introduce a simple penal
ization scheme that encourages Q-labels used across training batches to remain (
jointly) consistent with the expressible policy class. We also propose a search
framework that allows multiple Q-approximators to be generated and tracked, thus
mitigating the effect of premature (implicit) policy commitments. Experimental
results demonstrate that these methods can improve the performance of Q-learning
in a variety of Atari games, sometimes dramatically.
****

Where is the Information in a Deep Network?

Alessandro Achille,Stefano Soatto

Whatever information a deep neural network has gleaned from past data is encoded
in its weights. How this information affects the response of the network to fu
ture data is largely an open question. In fact, even how to define and measure
information in a network entails some subtleties. We measure information in the
weights of a deep neural network as the optimal trade-off between accuracy of th
e network and complexity of the weights relative to a prior. Depending on the p
rior, the definition reduces to known information measures such as Shannon Mutua
l Information and Fisher Information, but in general it affords added flexibilit
y that enables us to relate it to generalization, via the PAC-Bayes bound, and t
o invariance. For the latter, we introduce a notion of effective information in
the activations, which are deterministic functions of future inputs. We relate
this to the Information in the Weights, and use this result to show that models
of low (information) complexity not only generalize better, but are bound to le
arn invariant representations of future inputs. These relations hinge not only o
n the architecture of the model, but also on how it is trained.
****

Extreme Values are Accurate and Robust in Deep Networks

Jianguo Li,Mingjie Sun,Changshui Zhang

Recent evidence shows that convolutional neural networks (CNNs) are biased towar
ds textures so that CNNs are non-robust to adversarial perturbations over textur
es, while traditional robust visual features like SIFT (scale-invariant feature
transforms) are designed to be robust across a substantial range of affine disto
rtion, addition of noise, etc with the mimic of human perception nature. This pa
per aims to leverage good properties of SIFT to renovate CNN architectures towar
ds better accuracy and robustness. We borrow the scale-space extreme value idea
from SIFT, and propose EVPNet (extreme value preserving network) which contains
three novel components to model the extreme values: (1) parametric differences o
f Gaussian (DoG) to extract extrema, (2) truncated ReLU to suppress non-stable e
xtrema and (3) projected normalization layer (PNL) to mimic PCA-SIFT like featur
e normalization. Experiments demonstrate that EVPNets can achieve similar or bet
ter accuracy than conventional CNNs, while achieving much better robustness on a
set of adversarial attacks (FGSM,PGD,etc) even without adversarial training.
****

Statistically Consistent Saliency Estimation

Emre Barut,Shunyan Luo

The use of deep learning for a wide range of data problems has increased the nee
d for understanding and diagnosing these models, and deep learning interpretatio
n techniques have become an essential tool for data analysts. Although numerous
model interpretation methods have been proposed in recent years, most of these p

rocedures are based on heuristics with little or no theoretical guarantees. In t
his work, we propose a statistical framework for saliency estimation for black b
ox computer vision models. We build a model-agnostic estimation procedure that i
s statistically consistent and passes the saliency checks of Adebayo et al. (201
8). Our method requires solving a linear program, whose solution can be efficien
tly computed in polynomial time. Through our theoretical analysis, we establish
an upper bound on the number of model evaluations needed to recover the region o
f importance with high probability, and build a new perturbation scheme for esti
mation of local gradients that is shown to be more efficient than the commonly u
sed random perturbation schemes. Validity of the new method is demonstrated thro
ugh sensitivity analysis.

****

Domain-Independent Dominance of Adaptive Methods

Pedro Savarese,David McAllester,Sudarshan Babu,Michael Maire

From a simplified analysis of adaptive methods, we derive AvaGrad, a new optimiz
er which outperforms SGD on vision tasks when its adaptability is properly tuned
. We observe that the power of our method is partially explained by a decoupling
 of learning rate and adaptability, greatly simplifying hyperparameter search. I
n light of this observation, we demonstrate that, against conventional wisdom, A
dam can also outperform SGD on vision tasks, as long as the coupling between its
 learning rate and adaptability is taken into account. In practice, AvaGrad matc
hes the best results, as measured by generalization accuracy, delivered by any e
xisting optimizer (SGD or adaptive) across image classification (CIFAR, ImageNet
) and character-level language modelling (Penn Treebank) tasks. This later obser
vation, alongside of AvaGrad's decoupling of hyperparameters, could make it the
preferred optimizer for deep learning, replacing both SGD and Adam.

****

Neural Networks for Principal Component Analysis: A New Loss Function Provably Y
ields Ordered Exact Eigenvectors

Reza Oftadeh,Jiayi Shen,Zhangyang Wang,Dylan Shell

In this paper, we propose a new loss function for performing principal component
 analysis (PCA) using linear autoencoders (LAEs). Optimizing the standard L2 los
s results in a decoder matrix that spans the principal subspace of the sample co
variance of the data, but fails to identify the exact eigenvectors. This downsid
e originates from an invariance that cancels out in the global map. Here, we pro
ve that our loss function eliminates this issue, i.e. the decoder converges to t
he exact ordered unnormalized eigenvectors of the sample covariance matrix. For
this new loss, we establish that all local minima are global optima and also sho
w that computing the new loss (and also its gradients) has the same order of com
plexity as the classical loss. We report numerical results on both synthetic sim
ulations, and a real-data PCA experiment on MNIST (i.e., a 60,000 x784 matrix),
demonstrating our approach to be practically applicable and rectify previous LAE
s' downsides.

****

Symplectic ODE-Net: Learning Hamiltonian Dynamics with Control

Yaofeng Desmond Zhong,Biswadip Dey,Amit Chakraborty

In this paper, we introduce Symplectic ODE-Net (SymODEN), a deep learning framew
ork which can infer the dynamics of a physical system, given by an ordinary diff
erential equation (ODE), from observed state trajectories. To achieve better gen
eralization with fewer training samples, SymODEN incorporates appropriate induct
ive bias by designing the associated computation graph in a physics-informed man
ner. In particular, we enforce Hamiltonian dynamics with control to learn the un
derlying dynamics in a transparent way, which can then be leveraged to draw insi
ght about relevant physical aspects of the system, such as mass and potential en
ergy. In addition, we propose a parametrization which can enforce this Hamiltoni
an formalism even when the generalized coordinate data is embedded in a high-dim
ensional space or we can only access velocity data instead of generalized moment
um. This framework, by offering interpretable, physically-consistent models for
physical systems, opens up new possibilities for synthesizing model-based contro

l strategies.
****
Interpretations are useful: penalizing explanations to align neural networks wit h prior knowledge
Laura Rieger,Chandan Singh,W. James Murdoch,Bin Yu
For an explanation of a deep learning model to be effective, it must provide bot h insight into a model and suggest a corresponding action in order to achieve so me objective. Too often, the litany of proposed explainable deep learning metho ds stop at the first step, providing practitioners with insight into a model, but no way to act on it. In this paper, we propose contextual decomp osition explanation penalization (CDEP), a method which enables practitioners to leverage existing explanation methods in order to increase the predictive accur acy of deep learning models. In particular, when shown that a model has incorre ctly assigned importance to some features, CDEP enables practitioners to correct these errors by directly regularizing the provided explanations. Using explana tions provided by contextual decomposition (CD) (Murdoch et al., 2018), we demon strate the ability of our method to increase performance on an array of toy and real datasets.
****
FreeLB: Enhanced Adversarial Training for Natural Language Understanding
Chen Zhu,Yu Cheng,Zhe Gan,Siqi Sun,Tom Goldstein,Jingjing Liu
Adversarial training, which minimizes the maximal risk for label-preserving inpu t perturbations, has proved to be effective for improving the generalization of language models. In this work, we propose a novel adversarial training algorithm , FreeLB, that promotes higher invariance in the embedding space, by adding adve rsarial perturbations to word embeddings and minimizing the resultant adversaria l risk inside different regions around input samples. To validate the effectiven ess of the proposed approach, we apply it to Transformer-based models for natura l language understanding and commonsense reasoning tasks. Experiments on the GLU E benchmark show that when applied only to the finetuning stage, it is able to i mprove the overall test scores of BERT-base model from 78.3 to 79.4, and RoBERTa -large model from 88.5 to 88.8. In addition, the proposed approach achieves stat e-of-the-art single-model test accuracies of 85.44% and 67.75% on ARC-Easy and A RC-Challenge. Experiments on CommonsenseQA benchmark further demonstrate that Fr eeLB can be generalized and boost the performance of RoBERTa-large model on othe r tasks as well.
****
Behaviour Suite for Reinforcement Learning
Ian Osband,Yotam Doron,Matteo Hessel,John Aslanides,Eren Sezener,Andre Saraiva,K atrina McKinney,Tor Lattimore,Csaba Szepesvari,Satinder Singh,Benjamin Van Roy,R ichard Sutton,David Silver,Hado Van Hasselt
This paper introduces the Behaviour Suite for Reinforcement Learning, or bsuite for short. bsuite is a collection of carefully-designed experiments that investi gate core capabilities of reinforcement learning (RL) agents with two objectives . First, to collect clear, informative and scalable problems that capture key is sues in the design of general and efficient learning algorithms. Second, to stud y agent behaviour through their performance on these shared benchmarks. To compl ement this effort, we open source this http URL, which automates evaluation and analysis of any agent on bsuite. This library facilitates reproducible and acces sible research on the core issues in RL, and ultimately the design of superior l earning algorithms. Our code is Python, and easy to use within existing projects . We include examples with OpenAI Baselines, Dopamine as well as new reference i mplementations. Going forward, we hope to incorporate more excellent experiments from the research community, and commit to a periodic review of bsuite from a c ommittee of prominent researchers.
****
Strategies for Pre-training Graph Neural Networks
Weihua Hu*,Bowen Liu*,Joseph Gomes,Marinka Zitnik,Percy Liang,Vijay Pande,Jure L eskovec
Many applications of machine learning require a model to make accurate pre-dicti

ons on test examples that are distributionally different from training ones, whi
le task-specific labels are scarce during training. An effective approach to thi
s challenge is to pre-train a model on related tasks where data is abundant, and
 then fine-tune it on a downstream task of interest. While pre-training has been
 effective in many language and vision domains, it remains an open question how
to effectively use pre-training on graph datasets. In this paper, we develop a n
ew strategy and self-supervised methods for pre-training Graph Neural Networks (
GNNs). The key to the success of our strategy is to pre-train an expressive GNN
at the level of individual nodes as well as entire graphs so that the GNN can le
arn useful local and global representations simultaneously. We systematically st
udy pre-training on multiple graph classification datasets. We find that naïve s
trategies, which pre-train GNNs at the level of either entire graphs or individu
al nodes, give limited improvement and can even lead to negative transfer on man
y downstream tasks. In contrast, our strategy avoids negative transfer and impro
ves generalization significantly across downstream tasks, leading up to 9.4% abs
olute improvements in ROC-AUC over non-pre-trained models and achieving state-of
-the-art performance for molecular property prediction and protein function pred
iction.
****

GRAPHS, ENTITIES, AND STEP MIXTURE

Kyuyong Shin,Wonyoung Shin,Jung-Woo Ha,Sunyoung Kwon

Graph neural networks have shown promising results on representing and analyzing
 diverse graph-structured data such as social, citation, and protein interaction
 networks. Existing approaches commonly suffer from the oversmoothing issue, reg
ardless of whether policies are edge-based or node-based for neighborhood aggreg
ation. Most methods also focus on transductive scenarios for fixed graphs, leadi
ng to poor generalization performance for unseen graphs. To address these issues
, we propose a new graph neural network model that considers both edge-based nei
ghborhood relationships and node-based entity features, i.e. Graph Entities with
 Step Mixture via random walk (GESM). GESM employs a mixture of various steps th
rough random walk to alleviate the oversmoothing problem and attention to use no
de information explicitly. These two mechanisms allow for a weighted neighborhoo
d aggregation which considers the properties of entities and relations. With int
ensive experiments, we show that the proposed GESM achieves state-of-the-art or
comparable performances on four benchmark graph datasets comprising transductive
 and inductive learning tasks. Furthermore, we empirically demonstrate the signi
ficance of considering global information. The source code will be publicly avai
lable in the near future.
****

Refining the variational posterior through iterative optimization

Marton Havasi,Jasper Snoek,Dustin Tran,Jonathan Gordon,José Miguel Hernández-Lob
ato

Variational inference (VI) is a popular approach for approximate Bayesian infere
nce that is particularly promising for highly parameterized models such as deep
neural networks. A key challenge of variational inference is to approximate the
 posterior over model parameters with a distribution that is simpler and tractab
le yet sufficiently expressive. In this work, we propose a method for training h
ighly flexible variational distributions by starting with a coarse approximation
 and iteratively refining it. Each refinement step makes cheap, local adjustment
s and only requires optimization of simple variational families. We demonstrate
theoretically that our method always improves a bound on the approximation (the
Evidence Lower BOund) and observe this empirically across a variety of benchmark
 tasks. In experiments, our method consistently outperforms recent variational
inference methods for deep learning in terms of log-likelihood and the ELBO. We
 see that the gains are further amplified on larger scale models, significantly
outperforming standard VI and deep ensembles on residual networks on CIFAR10.
****

Aggregating explanation methods for neural networks stabilizes explanations

Laura Rieger,Lars Kai Hansen

■Despite a growing literature on explaining neural networks, no consensus has be

en reached on how to explain a neural network decision or how to evaluate an exp lanation.
■Our contributions in this paper are twofold. First, we investigate schemes to c ombine explanation methods and reduce model uncertainty to obtain a single aggre gated explanation. The aggregation is more robust and aligns better with the neu ral network than any single explanation method..
■Second, we propose a new approach to evaluating explanation methods that circum vents the need for manual evaluation and is not reliant on the alignment of neur al networks and humans decision processes.
****

Recurrent Hierarchical Topic-Guided Neural Language Models
Dandan Guo,Bo Chen,Ruiying Lu,Mingyuan Zhou
To simultaneously capture syntax and semantics from a text corpus, we propose a new larger-context language model that extracts recurrent hierarchical semantic structure via a dynamic deep topic model to guide natural language generation. M oving beyond a conventional language model that ignores long-range word dependen cies and sentence order, the proposed model captures not only intra-sentence wor d dependencies, but also temporal transitions between sentences and inter-senten ce topic dependences. For inference, we develop a hybrid of stochastic-gradient MCMC and recurrent autoencoding variational Bayes. Experimental results on a var iety of real-world text corpora demonstrate that the proposed model not only out performs state-of-the-art larger-context language models, but also learns interp retable recurrent multilayer topics and generates diverse sentences and paragrap hs that are syntactically correct and semantically coherent.
****

Invertible generative models for inverse problems: mitigating representation er ror and dataset bias
Muhammad Asim,Ali Ahmed,Paul Hand
Trained generative models have shown remarkable performance as priors for invers e problems in imaging. For example, Generative Adversarial Network priors permi t recovery of test images from 5-10x fewer measurements than sparsity priors. U nfortunately, these models may be unable to represent any particular image becau se of architectural choices, mode collapse, and bias in the training dataset. In this paper, we demonstrate that invertible neural networks, which have zero rep resentation error by design, can be effective natural signal priors at inverse p roblems such as denoising, compressive sensing, and inpainting. Our formulatio n is an empirical risk minimization that does not directly optimize the likeliho od of images, as one would expect. Instead we optimize the likelihood of the la tent representation of images as a proxy, as this is empirically easier.
For compressive sensing, our formulation can yield higher accuracy than sparsity priors across almost all undersampling ratios. For the same accuracy on test i mages, they can use 10-20x fewer measurements. We demonstrate that invertible p riors can yield better reconstructions than sparsity priors for images that have rare features of variation within the biased training set, including out-of-dis tribution natural images.
****

NAS-Bench-201: Extending the Scope of Reproducible Neural Architecture Search
Xuanyi Dong,Yi Yang
Neural architecture search (NAS) has achieved breakthrough success in a great nu mber of applications in the past few years.
It could be time to take a step back and analyze the good and bad aspects in the field of NAS. A variety of algorithms search architectures under different sear ch space. These searched architectures are trained using different setups, e.g., hyper-parameters, data augmentation, regularization. This raises a comparabilit y problem when comparing the performance of various NAS algorithms. NAS-Bench-10 1 has shown success to alleviate this problem. In this work, we propose an exten sion to NAS-Bench-101: NAS-Bench-201 with a different search space, results on m ultiple datasets, and more diagnostic information. NAS-Bench-201 has a fixed sea rch space and provides a unified benchmark for almost any up-to-date NAS algorit hms. The design of our search space is inspired by the one used in the most popu

lar cell-based searching algorithms, where a cell is represented as a directed a cyclic graph. Each edge here is associated with an operation selected from a pre defined operation set. For it to be applicable for all NAS algorithms, the searc h space defined in NAS-Bench-201 includes all possible architectures generated b y 4 nodes and 5 associated operation options, which results in 15,625 neural cel l candidates in total. The training log using the same setup and the performance for each architecture candidate are provided for three datasets. This allows re searchers to avoid unnecessary repetitive training for selected architecture and focus solely on the search algorithm itself. The training time saved for every architecture also largely improves the efficiency of most NAS algorithms and pre sents a more computational cost friendly NAS community for a broader range of re searchers. We provide additional diagnostic information such as fine-grained los s and accuracy, which can give inspirations to new designs of NAS algorithms. In further support of the proposed NAS-Bench-102, we have analyzed it from many as pects and benchmarked 10 recent NAS algorithms, which verify its applicability.
****

Learning World Graph Decompositions To Accelerate Reinforcement Learning
Wenling Shang,Alex Trott,Stephan Zheng,Caiming Xiong,Richard Socher
Efficiently learning to solve tasks in complex environments is a key challenge f or reinforcement learning (RL) agents. We propose to decompose a complex enviro nment using a task-agnostic world graphs, an abstraction that accelerates learni ng by enabling agents to focus exploration on a subspace of the environment.The nodes of a world graph are important waypoint states and edges represent feasibl e traversals between them. Our framework has two learning phases: 1) identifyin g world graph nodes and edges by training a binary recurrent variational auto-en coder (VAE) on trajectory data and 2) a hierarchical RL framework that leverages structural and connectivity knowledge from the learned world graph to bias expl oration towards task-relevant waypoints and regions. We show that our approach s ignificantly accelerates RL on a suite of challenging 2D grid world tasks: compa red to baselines, world graph integration doubles achieved rewards on simpler ta sks, e.g. MultiGoal, and manages to solve more challenging tasks, e.g. Door-Key, where baselines fail.
****

Gram-Gauss-Newton Method: Learning Overparameterized Neural Networks for Regress ion Problems
Tianle Cai*,Ruiqi Gao*,Jikai Hou*,Siyu Chen,Dong Wang,Di He,Zhihua Zhang,Liwei W ang
First-order methods such as stochastic gradient descent (SGD) are currently the standard algorithm for training deep neural networks. Second-order methods, desp ite their better convergence rate, are rarely used in practice due to the pro- h ibitive computational cost in calculating the second-order information. In this paper, we propose a novel Gram-Gauss-Newton (GGN) algorithm to train deep neural networks for regression problems with square loss. Our method draws inspiration from the connection between neural network optimization and kernel regression o f neural tangent kernel (NTK). Different from typical second-order methods that have heavy computational cost in each iteration, GGN only has minor overhead com pared to first-order methods such as SGD. We also give theoretical results to sh ow that for sufficiently wide neural networks, the convergence rate of GGN is qu adratic. Furthermore, we provide convergence guarantee for mini-batch GGN algori thm, which is, to our knowledge, the first convergence result for the mini-batch version of a second-order method on overparameterized neural net- works. Prelim inary experiments on regression tasks demonstrate that for training standard net works, our GGN algorithm converges much faster and achieves better performance t han SGD.
****

Controlling generative models with continuous factors of variations
Antoine Plumerault,Hervé Le Borgne,Céline Hudelot
Recent deep generative models can provide photo-realistic images as well as visu al or textual content embeddings useful to address various tasks of computer vis ion and natural language processing. Their usefulness is nevertheless often limi

ted by the lack of control over the generative process or the poor understanding of the learned representation. To overcome these major issues, very recent works have shown the interest of studying the semantics of the latent space of generative models. In this paper, we propose to advance on the interpretability of the latent space of generative models by introducing a new method to find meaningful directions in the latent space of any generative model along which we can move to control precisely specific properties of the generated image like position or scale of the object in the image. Our method is weakly supervised and particularly well suited for the search of directions encoding simple transformations of the generated image, such as translation, zoom or color variations. We demonstrate the effectiveness of our method qualitatively and quantitatively, both for GANs and variational auto-encoders.
****

Emergent Tool Use From Multi-Agent Autocurricula
Bowen Baker,Ingmar Kanitscheider,Todor Markov,Yi Wu,Glenn Powell,Bob McGrew,Igor Mordatch
Through multi-agent competition, the simple objective of hide-and-seek, and standard reinforcement learning algorithms at scale, we find that agents create a self-supervised autocurriculum inducing multiple distinct rounds of emergent strategy, many of which require sophisticated tool use and coordination. We find clear evidence of six emergent phases in agent strategy in our environment, each of which creates a new pressure for the opposing team to adapt; for instance, agents learn to build multi-object shelters using moveable boxes which in turn leads to agents discovering that they can overcome obstacles using ramps. We further provide evidence that multi-agent competition may scale better with increasing environment complexity and leads to behavior that centers around far more human-relevant skills than other self-supervised reinforcement learning methods such as intrinsic motivation. Finally, we propose transfer and fine-tuning as a way to quantitatively evaluate targeted capabilities, and we compare hide-and-seek agents to both intrinsic motivation and random initialization baselines in a suite of domain-specific intelligence tests.
****

The fairness-accuracy landscape of neural classifiers
Susan Wei,Marc Niethammer
That machine learning algorithms can demonstrate bias is well-documented by now. This work confronts the challenge of bias mitigation in feedforward fully-connected neural nets from the lens of causal inference and multiobjective optimisation. Regarding the former, a new causal notion of fairness is introduced that is particularly suited to giving a nuanced treatment of datasets collected under unfair practices. In particular, special attention is paid to subjects whose covariates could appear with substantial probability in either value of the sensitive attribute. Next, recognising that fairness and accuracy are competing objectives, the proposed methodology uses techniques from multiobjective optimisation to ascertain the fairness-accuracy landscape of a neural net classifier. Experimental results suggest that the proposed method produces neural net classifiers that distribute evenly across the Pareto front of the fairness-accuracy space and is more efficient at finding non-dominated points than an adversarial approach.
****

Simple and Effective Regularization Methods for Training on Noisily Labeled Data with Generalization Guarantee
Wei Hu,Zhiyuan Li,Dingli Yu
Over-parameterized deep neural networks trained by simple first-order methods are known to be able to fit any labeling of data. Such over-fitting ability hinders generalization when mislabeled training examples are present. On the other hand, simple regularization methods like early-stopping can often achieve highly nontrivial performance on clean test data in these scenarios, a phenomenon not theoretically understood. This paper proposes and analyzes two simple and intuitive regularization methods: (i) regularization by the distance between the network parameters to initialization, and (ii) adding a trainable auxiliary variable to the network output for each training example. Theoretically, we prove that gradi

ent descent training with either of these two methods leads to a generalization guarantee on the clean data distribution despite being trained using noisy labels. Our generalization analysis relies on the connection between wide neural network and neural tangent kernel (NTK). The generalization bound is independent of the network size, and is comparable to the bound one can get when there is no label noise. Experimental results verify the effectiveness of these methods on noisily labeled datasets.

****

Unsupervised Clustering using Pseudo-semi-supervised Learning

Divam Gupta,Ramachandran Ramjee,Nipun Kwatra,Muthian Sivathanu

In this paper, we propose a framework that leverages semi-supervised models to improve unsupervised clustering performance. To leverage semi-supervised models, we first need to automatically generate labels, called pseudo-labels. We find that prior approaches for generating pseudo-labels hurt clustering performance because of their low accuracy. Instead, we use an ensemble of deep networks to construct a similarity graph, from which we extract high accuracy pseudo-labels. The approach of finding high quality pseudo-labels using ensembles and training the semi-supervised model is iterated, yielding continued improvement. We show that our approach outperforms state of the art clustering results for multiple image and text datasets. For example, we achieve 54.6% accuracy for CIFAR-10 and 43.9% for 20news, outperforming state of the art by 8-12% in absolute terms.

****

Geometric Analysis of Nonconvex Optimization Landscapes for Overcomplete Learning

Qing Qu,Yuexiang Zhai,Xiao Li,Yuqian Zhang,Zhihui Zhu

Learning overcomplete representations finds many applications in machine learning and data analytics. In the past decade, despite the empirical success of heuristic methods, theoretical understandings and explanations of these algorithms are still far from satisfactory. In this work, we provide new theoretical insights for several important representation learning problems: learning (i) sparsely used overcomplete dictionaries and (ii) convolutional dictionaries. We formulate these problems as ℓ^4-norm optimization problems over the sphere and study the geometric properties of their nonconvex optimization landscapes. For both problems, we show the nonconvex objective has benign (global) geometric structures, which enable the development of efficient optimization methods finding the target solutions. Finally, our theoretical results are justified by numerical simulations.

****

PairNorm: Tackling Oversmoothing in GNNs

Lingxiao Zhao,Leman Akoglu

The performance of graph neural nets (GNNs) is known to gradually decrease with increasing number of layers. This decay is partly attributed to oversmoothing, where repeated graph convolutions eventually make node embeddings indistinguishable. We take a closer look at two different interpretations, aiming to quantify oversmoothing. Our main contribution is PairNorm, a novel normalization layer that is based on a careful analysis of the graph convolution operator, which prevents all node embeddings from becoming too similar. What is more, PairNorm is fast, easy to implement without any change to network architecture nor any additional parameters, and is broadly applicable to any GNN. Experiments on real-world graphs demonstrate that PairNorm makes deeper GCN, GAT, and SGC models more robust against oversmoothing, and significantly boosts performance for a new problem setting that benefits from deeper GNNs. Code is available at https://github.com/LingxiaoShawn/PairNorm.

****

Black-box Off-policy Estimation for Infinite-Horizon Reinforcement Learning

Ali Mousavi,Lihong Li,Qiang Liu,Denny Zhou

Off-policy estimation for long-horizon problems is important in many real-life applications such as healthcare and robotics, where high-fidelity simulators may not be available and on-policy evaluation is expensive or impossible. Recently,

\citet{liu18breaking} proposed an approach that avoids the curse of horizon suffered by typical importance-sampling-based methods. While showing promising results, this approach is limited in practice as it requires data being collected by a known behavior policy. In this work, we propose a novel approach that eliminates such limitations. In particular, we formulate the problem as solving for the fixed point of a "backward flow" operator and show that the fixed point solution gives the desired importance ratios of stationary distributions between the target and behavior policies. We analyze its asymptotic consistency and finite-sample
generalization. Experiments on benchmarks verify the effectiveness of our proposed approach.

****
Empirical Studies on the Properties of Linear Regions in Deep Neural Networks
Xiao Zhang,Dongrui Wu
A deep neural networks (DNN) with piecewise linear activations can partition the input space into numerous small linear regions, where different linear functions are fitted. It is believed that the number of these regions represents the expressivity of a DNN. This paper provides a novel and meticulous perspective to look into DNNs: Instead of just counting the number of the linear regions, we study their local properties, such as the inspheres, the directions of the corresponding hyperplanes, the decision boundaries, and the relevance of the surrounding regions. We empirically observed that different optimization techniques lead to completely different linear regions, even though they result in similar classification accuracies. We hope our study can inspire the design of novel optimization techniques, and help discover and analyze the behaviors of DNNs.
****
SNOW: Subscribing to Knowledge via Channel Pooling for Transfer & Lifelong Learning of Convolutional Neural Networks
Chungkuk Yoo,Bumsoo Kang,Minsik Cho
SNOW is an efficient learning method to improve training/serving throughput as well as accuracy for transfer and lifelong learning of convolutional neural networks based on knowledge subscription. SNOW selects the top-K useful intermediate feature maps for a target task from a pre-trained and frozen source model through a novel channel pooling scheme, and utilizes them in the task-specific delta model. The source model is responsible for generating a large number of generic feature maps. Meanwhile, the delta model selectively subscribes to those feature maps and fuses them with its local ones to deliver high accuracy for the target task. Since a source model takes part in both training and serving of all target tasks
in an inference-only mode, one source model can serve multiple delta models, enabling significant computation sharing. The sizes of such delta models are fractional of the source model, thus SNOW also provides model-size efficiency.
Our experimental results show that SNOW offers a superior balance between accuracy and training/inference speed for various image classification tasks to the existing transfer and lifelong learning practices.
****
Smoothness and Stability in GANs
Casey Chu,Kentaro Minami,Kenji Fukumizu
Generative adversarial networks, or GANs, commonly display unstable behavior during training. In this work, we develop a principled theoretical framework for understanding the stability of various types of GANs. In particular, we derive conditions that guarantee eventual stationarity of the generator when it is trained with gradient descent, conditions that must be satisfied by the divergence that is minimized by the GAN and the generator's architecture. We find that existing GAN variants satisfy some, but not all, of these conditions. Using tools from convex analysis, optimal transport, and reproducing kernels, we construct a GAN that fulfills these conditions simultaneously. In the process, we explain and clarify the need for various existing GAN stabilization techniques, including Lipschitz constraints, gradient penalties, and smooth activation functions.

****
Adaptive Correlated Monte Carlo for Contextual Categorical Sequence Generation
Xinjie Fan,Yizhe Zhang,Zhendong Wang,Mingyuan Zhou
Sequence generation models are commonly refined with reinforcement learning over user-defined metrics. However, high gradient variance hinders the practical use of this method. To stabilize this method, we adapt to contextual generation of categorical sequences a policy gradient estimator, which evaluates a set of correlated Monte Carlo (MC) rollouts for variance control. Due to the correlation, the number of unique rollouts is random and adaptive to model uncertainty; those rollouts naturally become baselines for each other, and hence are combined to effectively reduce gradient variance. We also demonstrate the use of correlated MC rollouts for binary-tree softmax models, which reduce the high generation cost in large vocabulary scenarios by decomposing each categorical action into a sequence of binary actions. We evaluate our methods on both neural program synthesis and image captioning. The proposed methods yield lower gradient variance and consistent improvement over related baselines.
****
On Bonus Based Exploration Methods In The Arcade Learning Environment
Adrien Ali Taiga,William Fedus,Marlos C. Machado,Aaron Courville,Marc G. Bellemare
Research on exploration in reinforcement learning, as applied to Atari 2600 game-playing, has emphasized tackling difficult exploration problems such as Montezuma's Revenge (Bellemare et al., 2016). Recently, bonus-based exploration methods, which explore by augmenting the environment reward, have reached above-human average performance on such domains. In this paper we reassess popular bonus-based exploration methods within a common evaluation framework. We combine Rainbow (Hessel et al., 2018) with different exploration bonuses and evaluate its performance on Montezuma's Revenge, Bellemare et al.'s set of hard of exploration games with sparse rewards, and the whole Atari 2600 suite. We find that while exploration bonuses lead to higher score on Montezuma's Revenge they do not provide meaningful gains over the simpler epsilon-greedy scheme. In fact, we find that methods that perform best on that game often underperform epsilon-greedy on easy exploration Atari 2600 games. We find that our conclusions remain valid even when hyperparameters are tuned for these easy-exploration games. Finally, we find that none of the methods surveyed benefit from additional training samples (1 billion frames, versus Rainbow's 200 million) on Bellemare et al.'s hard exploration games. Our results suggest that recent gains in Montezuma's Revenge may be better attributed to architecture change, rather than better exploration schemes; and that the real pace of progress in exploration research for Atari 2600 games may have been obfuscated by good results on a single domain.
****
Power up! Robust Graph Convolutional Network based on Graph Powering
Ming Jin,Heng Chang,Wenwu Zhu,Somayeh Sojoudi
Graph convolutional networks (GCNs) are powerful tools for graph-structured data. However, they have been recently shown to be vulnerable to topological attacks. To enhance adversarial robustness, we go beyond spectral graph theory to robust graph theory. By challenging the classical graph Laplacian, we propose a new convolution operator that is provably robust in the spectral domain and is incorporated in the GCN architecture to improve expressivity and interpretability. By extending the original graph to a sequence of graphs, we also propose a robust training paradigm that encourages transferability across graphs that span a range of spatial and spectral characteristics. The proposed approaches are demonstrated in extensive experiments to {simultaneously} improve performance in both benign and adversarial situations.
****
Global graph curvature
Liudmila Prokhorenkova,Egor Samosvat,Pim van der Hoorn
Recently, non-Euclidean spaces became popular for embedding structured data. However, determining suitable geometry and, in particular, curvature for a given dataset is still an open problem. In this paper, we define a notion of global grap

h curvature, specifically catered to the problem of embedding graphs, and analyz
e the problem of estimating this curvature using only graph-based characteristic
s (without actual graph embedding). We show that optimal curvature essentially d
epends on dimensionality of the embedding space and loss function one aims to mi
nimize via embedding. We review the existing notions of local curvature (e.g., O
llivier-Ricci curvature) and analyze their properties theoretically and empirica
lly. In particular, we show that such curvatures are often unable to properly es
timate the global one. Hence, we propose a new estimator of global graph curvatu
re specifically designed for zero-one loss function.
****

Deep k-NN for Noisy Labels
Dara Bahri,Heinrich Jiang,Maya Gupta
Modern machine learning models are often trained on examples with noisy labels t
hat hurt performance and are hard to identify. In this paper, we provide an empi
rical study showing that a simple k-nearest neighbor-based filtering approach
on the logit layer of a preliminary model can remove mislabeled training data a
nd produce more accurate models than some recently proposed methods. We also pro
vide new statistical guarantees into its efficacy.
****

Filling the Soap Bubbles: Efficient Black-Box Adversarial Certification with Non
-Gaussian Smoothing
Dinghuai Zhang*,Mao Ye*,Chengyue Gong*,Zhanxing Zhu,Qiang Liu
Randomized classifiers have been shown to provide a promising approach for achie
ving certified robustness against adversarial attacks in deep learning. However,
 most existing methods only leverage Gaussian smoothing noise and only work for
ℓ_2 perturbation. We propose a general framework of adversarial certificati
on with non-Gaussian noise and for more general types of attacks, from a unified
 functional optimization perspective. Our new framework allows us to identify a
key trade-off between accuracy and robustness via designing smoothing distributi
ons, helping to design two new families of non-Gaussian smoothing distributions
that work more efficiently for ℓ_2 and ℓ_∞ attacks, respectively.
 Our proposed methods achieve better results than previous works and provide a n
ew perspective on randomized smoothing certification.
****

Guided Adaptive Credit Assignment for Sample Efficient Policy Optimization
Hao Liu,Richard Socher,Caiming Xiong
Policy gradient methods have achieved remarkable successes in solving challengin
g reinforcement learning problems. However, it still often suffers from sparse r
eward tasks, which leads to poor sample efficiency during training. In this work
, we propose a guided adaptive credit assignment method to do effectively credit
 assignment for policy gradient methods. Motivated by entropy regularized policy
 optimization, our method extends the previous credit assignment methods by intr
oducing more general guided adaptive credit assignment(GACA). The benefit of GAC
A is a principled way of utilizing off-policy samples. The effectiveness of prop
osed algorithm is demonstrated on the challenging \textsc{WikiTableQuestions} an
d \textsc{WikiSQL} benchmarks and an instruction following environment. The task
 is generating action sequences or program sequences from natural language quest
ions or instructions, where only final binary success-failure execution feedback
 is available. Empirical studies show that our method significantly improves the
 sample efficiency of the state-of-the-art policy optimization approaches.
****

A Theory of Usable Information under Computational Constraints
Yilun Xu,Shengjia Zhao,Jiaming Song,Russell Stewart,Stefano Ermon
We propose a new framework for reasoning about information in complex systems. O
ur foundation is based on a variational extension of Shannon’s information theor
y that takes into account the modeling power and computational constraints of th
e observer. The resulting predictive V-information encompasses mutual informatio
n and other notions of informativeness such as the coefficient of determination.
 Unlike Shannon’s mutual information and in violation of the data processing ine
quality, V-information can be created through computation. This is consistent wi

th deep neural networks extracting hierarchies of progressively more informative features in representation learning. Additionally, we show that by incorporating computational constraints, V-information can be reliably estimated from data even in high dimensions with PAC-style guarantees. Empirically, we demonstrate predictive V-information is more effective than mutual information for structure learning and fair representation learning. Codes are available at https://github.com/Newbeeer/V-information .

****

On the Invertibility of Invertible Neural Networks

Jens Behrmann,Paul Vicol,Kuan-Chieh Wang,Roger B. Grosse,Jörn-Henrik Jacobsen

Guarantees in deep learning are hard to achieve due to the interplay of flexible modeling schemes and complex tasks. Invertible neural networks (INNs), however, provide several mathematical guarantees by design, such as the ability to approximate non-linear diffeomorphisms. One less studied advantage of INNs is that they enable the design of bi-Lipschitz functions. This property has been used implicitly by various works to design generative models, memory-saving gradient computation, regularize classifiers, and solve inverse problems.

In this work, we study Lipschitz constants of invertible architectures in order to investigate guarantees on stability of their inverse and forward mapping. Our analysis reveals that commonly-used INN building blocks can easily become non-invertible, leading to questionable ``exact'' log likelihood computations and training difficulties. We introduce a set of numerical analysis tools to diagnose non-invertibility in practice. Finally, based on our theoretical analysis, we show how to guarantee numerical invertibility for one of the most common INN architectures.

****

Shallow VAEs with RealNVP Prior Can Perform as Well as Deep Hierarchical VAEs

Haowen Xu,Wenxiao Chen,Jinlin Lai,Zhihan Li,Youjian Zhao,Dan Pei

Using powerful posterior distributions is a popular technique in variational inference. However, recent works showed that the aggregated posterior may fail to match unit Gaussian prior, even with expressive posteriors, thus learning the prior becomes an alternative way to improve the variational lower-bound. We show that using learned RealNVP prior and just one latent variable in VAE, we can achieve test NLL comparable to very deep state-of-the-art hierarchical VAE, outperforming many previous works with complex hierarchical VAE architectures. We hypothesize that, when coupled with Gaussian posteriors, the learned prior can encourage appropriate posterior overlapping, which is likely to improve reconstruction loss and lower-bound, supported by our experimental results. We demonstrate that, with learned RealNVP prior, ß-VAE can have better rate-distortion curve than using fixed Gaussian prior.

****

GAN-based Gaussian Mixture Model Responsibility Learning

Wanming Huang,Shuai Jiang,Xuan Liang,Ian Oppermann,Richard Yi Da Xu

Mixture Model (MM) is a probabilistic framework which allows us to define a data set containing K different modes. When each of the modes is associated with a Gaussian distribution, we refer it as Gaussian MM, or GMM. Given a data point x, GMM may assume the existence of a random index $k \in \{1, . . . , K\}$ identifying which Gaussian the particular data is associated with. In a traditional GMM paradigm, it is straightforward to compute in closed-form, the conditional like- lihood $p(x|k, \theta)$, as well as responsibility probability $p(k|x, \theta)$ which describes the distribution index corresponds to the data. Computing the responsibility allows us to retrieve many important statistics of the overall dataset, including the weights of each of the modes. Modern large datasets often contain multiple unlabelled modes, such as paintings dataset containing several styles; fashion images containing several unlabelled categories. In its raw representation, the Euclidean distances between the data do not allow them to form mixtures naturally, nor it’s feasible to compute responsibility distribution, making GMM unable to apply. To this paper, we utilize the Generative Adversarial Network (GAN) framework to achieve an alternative plausible method to compute these probabilities at the data’s latent space z instead of x. Instead of defining $p(x|k, \theta)$ explicitly, w

e devised a modified GAN to allow us to define the distribution using p(z|k, θ), where z is the corresponding latent representation of x, as well as p(k|x, θ) through an additional classification network which is trained with the GAN in an "end-to-end" fashion. These techniques allow us to discover interesting properties of an unsupervised dataset, including dataset segments as well as generating new "out-distribution" data by smooth linear interpolation across any combinations of the modes in a completely unsupervised manner.
****
Information-Theoretic Local Minima Characterization and Regularization
Zhiwei Jia,Hao Su
Recent advances in deep learning theory have evoked the study of generalizability across different local minima of deep neural networks (DNNs). While current work focused on either discovering properties of good local minima or developing regularization techniques to induce good local minima, no approach exists that can tackle both problems. We achieve these two goals successfully in a unified manner. Specifically, based on the Fisher information we propose a metric both strongly indicative of generalizability of local minima and effectively applied as a practical regularizer. We provide theoretical analysis including a generalization bound and empirically demonstrate the success of our approach in both capturing and improving the generalizability of DNNs. Experiments are performed on CIFAR-10 and CIFAR-100 for various network architectures.
****
Well-Read Students Learn Better: On the Importance of Pre-training Compact Models
Iulia Turc,Ming-Wei Chang,Kenton Lee,Kristina Toutanova
Recent developments in natural language representations have been accompanied by large and expensive models that leverage vast amounts of general-domain text through self-supervised pre-training. Due to the cost of applying such models to down-stream tasks, several model compression techniques on pre-trained language representations have been proposed (Sun et al., 2019; Sanh, 2019). However, surprisingly, the simple baseline of just pre-training and fine-tuning compact models has been overlooked. In this paper, we first show that pre-training remains important in the context of smaller architectures, and fine-tuning pre-trained compact models can be competitive to more elaborate methods proposed in concurrent work. Starting with pre-trained compact models, we then explore transferring task knowledge from large fine-tuned models through standard knowledge distillation. The resulting simple, yet effective and general algorithm, Pre-trained Distillation, brings further improvements. Through extensive experiments, we more generally explore the interaction between pre-training and distillation under two variables that have been under-studied: model size and properties of unlabeled task data. One surprising observation is that they have a compound effect even when sequentially applied on the same data. To accelerate future research, we will make our 24 pre-trained miniature BERT models publicly available.
****
IMPACT: Importance Weighted Asynchronous Architectures with Clipped Target Networks
Michael Luo,Jiahao Yao,Richard Liaw,Eric Liang,Ion Stoica
The practical usage of reinforcement learning agents is often bottlenecked by the duration of training time. To accelerate training, practitioners often turn to distributed reinforcement learning architectures to parallelize and accelerate the training process. However, modern methods for scalable reinforcement learning (RL) often tradeoff between the throughput of samples that an RL agent can learn from (sample throughput) and the quality of learning from each sample (sample efficiency). In these scalable RL architectures, as one increases sample throughput (i.e. increasing parallelization in IMPALA (Espeholt et al., 2018)), sample efficiency drops significantly. To address this, we propose a new distributed reinforcement learning algorithm, IMPACT. IMPACT extends PPO with three changes: a target network for stabilizing the surrogate objective, a circular buffer, and truncated importance sampling. In discrete action-space environments, we show that IMPACT attains higher reward and, simultaneously, achieves up to 30% decreas

e in training wall-time than that of IMPALA. For continuous control environments, IMPACT trains faster than existing scalable agents while preserving the sample efficiency of synchronous PPO.
****
UWGAN: UNDERWATER GAN FOR REAL-WORLD UNDERWATER COLOR RESTORATION AND DEHAZING
Nan Wang,Yabin Zhou,Fenglei Han,Lichao Wan,Haitao Zhu,Yaojing Zheng
In real-world underwater environment, exploration of seabed resources, underwater archaeology, and underwater fishing rely on a variety of sensors, vision sensor is the most important one due to its high information content, non-intrusive, and passive nature. However, wavelength-dependent light attenuation and back-scattering result in color distortion and haze effect, which degrade the visibility of images. To address this problem, firstly, we proposed an unsupervised generative adversarial network (GAN) for generating realistic underwater images (color distortion and haze effect simulation) from in-air image and depth map pairs. Secondly, U-Net, which is trained efficiently using synthetic underwater dataset, is adopted for color restoration and de-hazing. Our model directly reconstructs underwater clear images using end-to-end autoencoder networks, while maintaining scene content structural similarity. The results obtained by our method were compared with existing methods qualitatively and quantitatively. Experimental results on open real-world underwater datasets demonstrate that the presented method performs well on different actual underwater scenes, and the processing speed can reach up to 125FPS on images running on one NVIDIA 1060 GPU.
****
HiLLoC: lossless image compression with hierarchical latent variable models
James Townsend,Thomas Bird,Julius Kunze,David Barber
We make the following striking observation: fully convolutional VAE models trained on 32x32 ImageNet can generalize well, not just to 64x64 but also to far larger photographs, with no changes to the model. We use this property, applying fully convolutional models to lossless compression, demonstrating a method to scale the VAE-based 'Bits-Back with ANS' algorithm for lossless compression to large color photographs, and achieving state of the art for compression of full size ImageNet images. We release Craystack, an open source library for convenient prototyping of lossless compression using probabilistic models, along with full implementations of all of our compression results.
****
Learning to Learn Kernels with Variational Random Features
Haoliang Sun,Yingjun Du,Jun Xu,Yilong Yin,Xiantong Zhen,Ling Shao
Meta-learning for few-shot learning involves a meta-learner that acquires shared knowledge from a set of prior tasks to improve the performance of a base-learner on new tasks with a small amount of data. Kernels are commonly used in machine learning due to their strong nonlinear learning capacity, which have not yet been fully investigated in the meta-learning scenario for few-shot learning. In this work, we explore kernel approximation with random Fourier features in the meta-learning framework for few-shot learning. We propose learning adaptive kernels by meta variational random features (MetaVRF), which is formulated as a variational inference problem. To explore shared knowledge across diverse tasks, our MetaVRF deploys an LSTM inference network to generate informative features, which can establish kernels of highly representational power with low spectral sampling rates, while also being able to quickly adapt to specific tasks for improved performance. We evaluate MetaVRF on a variety of few-shot learning tasks for both regression and classification. Experimental results demonstrate that our MetaVRF can deliver much better or competitive performance than recent meta-learning algorithms.
****
Efficient Wrapper Feature Selection using Autoencoder and Model Based Elimination
Sharan Ramjee,Aly El Gamal
We propose a computationally efficient wrapper feature selection method - called Autoencoder and Model Based Elimination of features using Relevance and Redundancy scores (AMBER) - that uses a single ranker model along with autoencoders to

perform greedy backward elimination of features. The ranker model is used to pri
oritize the removal of features that are not critical to the classification task
, while the autoencoders are used to prioritize the elimination of correlated fe
atures. We demonstrate the superior feature selection ability of AMBER on 4 well
 known datasets corresponding to different domain applications via comparing the
 accuracies with other computationally efficient state-of-the-art feature select
ion techniques. Interestingly, we find that the ranker model that is used for fe
ature selection does not necessarily have to be the same as the final classifier
 that is trained on the selected features. Finally, we hypothesize that overfitt
ing the ranker model on the training set facilitates the selection of more salie
nt features.
****

Physics-aware Difference Graph Networks for Sparsely-Observed Dynamics
Sungyong Seo*,Chuizheng Meng*,Yan Liu
Sparsely available data points cause numerical error on finite differences which
 hinders us from modeling the dynamics of physical systems. The discretization e
rror becomes even larger when the sparse data are irregularly distributed or def
ined on an unstructured grid, making it hard to build deep learning models to ha
ndle physics-governing observations on the unstructured grid. In this paper, we
propose a novel architecture, Physics-aware Difference Graph Networks (PA-DGN),
which exploits neighboring information to learn finite differences inspired by p
hysics equations. PA-DGN leverages data-driven end-to-end learning to discover u
nderlying dynamical relations between the spatial and temporal differences in gi
ven sequential observations. We demonstrate the superiority of PA-DGN in the app
roximation of directional derivatives and the prediction of graph signals on the
 synthetic data and the real-world climate observations from weather stations.
****

Polylogarithmic width suffices for gradient descent to achieve arbitrarily small
 test error with shallow ReLU networks
Ziwei Ji,Matus Telgarsky
Recent theoretical work has guaranteed that overparameterized networks trained b
y gradient descent achieve arbitrarily low training error, and sometimes even lo
w test error.
The required width, however, is always polynomial in at least one of the sample
size n, the (inverse) target error $1/\epsilon$, and the (inverse) failure pro
bability $1/\delta$.
This work shows that $\widetilde{\Theta}(1/\epsilon)$ iterations of gradient des
cent with $\widetilde{\Omega}(1/\epsilon^2)$ training examples on two-layer ReLU
 networks of any width exceeding $\textrm{polylog}(n,1/\epsilon,1/\delta)$ suffi
ce to achieve a test misclassification error of ϵ.
We also prove that stochastic gradient descent can achieve ϵ test error
 with polylogarithmic width and $\widetilde{\Theta}(1/\epsilon)$ samples.
The analysis relies upon the separation margin of the limiting kernel, which is
guaranteed positive, can distinguish between true labels and random labels, and
can give a tight sample-complexity analysis in the infinite-width setting.
****

Enhancing Language Emergence through Empathy
Marie Ossenkopf
The emergence of language in multi-agent settings is a promising research direct
ion to ground natural language in simulated agents. If AI would be able to under
stand the meaning of language through its using it, it could also transfer it to
 other situations flexibly. That is seen as an important step towards achieving
general AI. The scope of emergent communication is so far, however, still limite
d. It is necessary to enhance the learning possibilities for skills associated w
ith communication to increase the emergable complexity. We took an example from
human language acquisition and the importance of the empathic connection in this
 process. We propose an approach to introduce the notion of empathy to multi-age
nt deep reinforcement learning. We extend existing approaches on referential gam
es with an auxiliary task for the speaker to predict the listener's mind change
improving the learning time. Our experiments show the high potential of this arc

hitectural element by doubling the learning speed of the test setup.
****
The Generalization-Stability Tradeoff in Neural Network Pruning
Brian R. Bartoldson,Ari S. Morcos,Adrian Barbu,Gordon Erlebacher
Pruning neural network parameters is often viewed as a means to compress models, but pruning has also been motivated by the desire to prevent overfitting. This motivation is particularly relevant given the perhaps surprising observation that a wide variety of pruning approaches increase test accuracy despite sometimes massive reductions in parameter counts. To better understand this phenomenon, we analyze the behavior of pruning over the course of training, finding that pruning's effect on generalization relies more on the instability it generates (defined as the drops in test accuracy immediately following pruning) than on the final size of the pruned model. We demonstrate that even the pruning of unimportant parameters can lead to such instability, and show similarities between pruning and regularizing by injecting noise, suggesting a mechanism for pruning-based generalization improvements that is compatible with the strong generalization recently observed in over-parameterized networks.
****
Word embedding re-examined: is the symmetrical factorization optimal?
Zhichao Han,Jia Li,Xu Li,Hong Cheng
As observed in previous works, many word embedding methods exhibit two interesting properties: (1) words having similar semantic meanings are embedded closely; (2) analogy structure exists in the embedding space, such that ''emph{Paris} is to \emph{France} as \emph{Berlin} is to \emph{Germany}''. We theoretically analyze the inner mechanism leading to these nice properties. Specifically, the embedding can be viewed as a linear transformation from the word-context co-occurrence space to the embedding space. We reveal how the relative distances between nodes change during this transforming process. Such linear transformation will result in these good properties. Based on the analysis, we also provide the answer to a question whether the symmetrical factorization (e.g., \texttt{word2vec}) is better than traditional SVD method. We propose a method to improve the embedding further. The experiments on real datasets verify our analysis.
****
Empowering Graph Representation Learning with Paired Training and Graph Co-Attention
Andreea Deac,Yu-Hsiang Huang,Petar Velickovic,Pietro Lio,Jian Tang
Through many recent advances in graph representation learning, performance achieved on tasks involving graph-structured data has substantially increased in recent years---mostly on tasks involving node-level predictions. The setup of prediction tasks over entire graphs (such as property prediction for a molecule, or side-effect prediction for a drug), however, proves to be more challenging, as the algorithm must combine evidence about several structurally relevant patches of the graph into a single prediction.
Most prior work attempts to predict these graph-level properties while considering only one graph at a time---not allowing the learner to directly leverage structural similarities and motifs across graphs. Here we propose a setup in which a graph neural network receives pairs of graphs at once, and extend it with a co-attentional layer that allows node representations to easily exchange structural information across them. We first show that such a setup provides natural benefits on a pairwise graph classification task (drug-drug interaction prediction), and then expand to a more generic graph regression setup: enhancing predictions over QM9, a standard molecular prediction benchmark. Our setup is flexible, powerful and makes no assumptions about the underlying dataset properties, beyond anticipating the existence of multiple training graphs.
****
Learning representations for binary-classification without backpropagation
Mathias Lechner
The family of feedback alignment (FA) algorithms aims to provide a more biologically motivated alternative to backpropagation (BP), by substituting the computations that are unrealistic to be implemented in physical brains.

While FA algorithms have been shown to work well in practice, there is a lack of rigorous theory proofing their learning capabilities.■■
Here we introduce the first feedback alignment algorithm with provable learning guarantees. In contrast to existing work, we do not require any assumption about the size or depth of the network except that it has a single output neuron, i.e., such as for binary classification tasks.
We show that our FA algorithm can deliver its theoretical promises in practice, surpassing the learning performance of existing FA methods and matching backpropagation in binary classification tasks.
Finally, we demonstrate the limits of our FA variant when the number of output neurons grows beyond a certain quantity.
****
Deep unsupervised feature selection
Ian Covert,Uygar Sumbul,Su-In Lee
Unsupervised feature selection involves finding a small number of highly informative features, in the absence of a specific supervised learning task. Selecting a small number of features is an important problem in many scientific domains with high-dimensional observations. Here, we propose the restricted autoencoder (RAE) framework for selecting features that can accurately reconstruct the rest of the features. We justify our approach through a novel proof that the reconstruction ability of a set of features bounds its performance in downstream supervised learning tasks. Based on this theory, we present a learning algorithm for RAEs that iteratively eliminates features using learned per-feature corruption rates. We apply the RAE framework to two high-dimensional biological datasets—single cell RNA sequencing and microarray gene expression data, which pose important problems in cell biology and precision medicine—and demonstrate that RAEs outperform nine baseline methods, often by a large margin.
****
WaveFlow: A Compact Flow-based Model for Raw Audio
Wei Ping,Kainan Peng,Kexin Zhao,Zhao Song
In this work, we present WaveFlow, a small-footprint generative flow for raw audio, which is trained with maximum likelihood without complicated density distillation and auxiliary losses as used in Parallel WaveNet. It provides a unified view of flow-based models for raw audio, including autoregressive flow (e.g., WaveNet) and bipartite flow (e.g., WaveGlow) as special cases. We systematically study these likelihood-based generative models for raw waveforms in terms of test likelihood and speech fidelity. We demonstrate that WaveFlow can synthesize high-fidelity speech and obtain comparable likelihood as WaveNet, while only requiring a few sequential steps to generate very long waveforms. In particular, our small-footprint WaveFlow has only 5.91M parameters and can generate 22.05kHz speech 15.39 times faster than real-time on a GPU without customized inference kernels.
****
Mathematical Reasoning in Latent Space
Dennis Lee,Christian Szegedy,Markus Rabe,Sarah Loos,Kshitij Bansal
We design and conduct a simple experiment to study whether neural networks can perform several steps of approximate reasoning in a fixed dimensional latent space. The set of rewrites (i.e. transformations) that can be successfully performed on a statement represents essential semantic features of the statement. We can compress this information by embedding the formula in a vector space, such that the vector associated with a statement can be used to predict whether a statement can be rewritten by other theorems. Predicting the embedding of a formula generated by some rewrite rule is naturally viewed as approximate reasoning in the latent space. In order to measure the effectiveness of this reasoning, we perform approximate deduction sequences in the latent space and use the resulting embedding to inform the semantic features of the corresponding formal statement (which is obtained by performing the corresponding rewrite sequence using real formulas). Our experiments show that graph neural networks can make non-trivial predictions about the rewrite-success of statements, even when they propagate predicted latent representations for several steps. Since our corpus of mathematical for

mulas includes a wide variety of mathematical disciplines, this experiment is a strong indicator for the feasibility of deduction in latent space in general.
****

Black Box Recursive Translations for Molecular Optimization
Farhan Damani,Vishnu Sresht,Stephen Ra
Machine learning algorithms for generating molecular structures offer a promising new approach to drug discovery. We cast molecular optimization as a translation problem, where the goal is to map an input compound to a target compound with improved biochemical properties. Remarkably, we observe that when generated molecules are iteratively fed back into the translator, molecular compound attributes improve with each step. We show that this finding is invariant to the choice of translation model, making this a "black box" algorithm. We call this method Black Box Recursive Translation (BBRT), a new inference method for molecular property optimization. This simple, powerful technique operates strictly on the inputs and outputs of any translation model. We obtain new state-of-the-art results for molecular property optimization tasks using our simple drop-in replacement with well-known sequence and graph-based models. Our method provides a significant boost in performance relative to its non-recursive peers with just a simple "``for" loop. Further, BBRT is highly interpretable, allowing users to map the evolution of newly discovered compounds from known starting points.
****

Improved Generalization Bound of Permutation Invariant Deep Neural Networks
Akiyoshi Sannai,Masaaki Imaizumi
We theoretically prove that a permutation invariant property of deep neural networks largely improves its generalization performance. Learning problems with data that are invariant to permutations are frequently observed in various applications, for example, point cloud data and graph neural networks. Numerous methodologies have been developed and they achieve great performances, however, understanding a mechanism of the performance is still a developing problem. In this paper, we derive a theoretical generalization bound for invariant deep neural networks with a ReLU activation to clarify their mechanism. Consequently, our bound shows that the main term of their generalization gap is improved by $\sqrt{n!}$ where n is a number of permuting coordinates of data. Moreover, we prove that an approximation power of invariant deep neural networks can achieve an optimal rate, though the networks are restricted to be invariant. To achieve the results, we develop several new proof techniques such as correspondence with a fundamental domain and a scale-sensitive metric entropy.
****

Frequency-based Search-control in Dyna
Yangchen Pan,Jincheng Mei,Amir-massoud Farahmand
Model-based reinforcement learning has been empirically demonstrated as a successful strategy to improve sample efficiency. In particular, Dyna is an elegant model-based architecture integrating learning and planning that provides huge flexibility of using a model. One of the most important components in Dyna is called search-control, which refers to the process of generating state or state-action pairs from which we query the model to acquire simulated experiences. Search-control is critical in improving learning efficiency. In this work, we propose a simple and novel search-control strategy by searching high frequency regions of the value function. Our main intuition is built on Shannon sampling theorem from signal processing, which indicates that a high frequency signal requires more samples to reconstruct. We empirically show that a high frequency function is more difficult to approximate. This suggests a search-control strategy: we should use states from high frequency regions of the value function to query the model to acquire more samples. We develop a simple strategy to locally measure the frequency of a function by gradient and hessian norms, and provide theoretical justification for this approach. We then apply our strategy to search-control in Dyna, and conduct experiments to show its property and effectiveness on benchmark domains.
****

Off-policy Bandits with Deficient Support

Noveen Sachdeva,Yi Su,Thorsten Joachims
Off-policy training of contextual-bandit policies is attractive in online systems (e.g. search, recommendation, ad placement), since it enables the reuse of large amounts of log data from the production system. State-of-the-art methods for off-policy learning, however, are based on inverse propensity score (IPS) weighting, which requires that the logging policy chooses all actions with non-zero probability for any context (i.e., full support). In real-world systems, this condition is often violated, and we show that existing off-policy learning methods based on IPS weighting can fail catastrophically. We therefore develop new off-policy contextual-bandit methods that can controllably and robustly learn even when the logging policy has deficient support. To this effect, we explore three approaches that provide various guarantees for safe learning despite the inherent limitations of support deficient data: restricting the action space, reward extrapolation, and restricting the policy space. We analyze the statistical and computational properties of these three approaches, and empirically evaluate their effectiveness in a series of experiments. We find that controlling the policy space is both computationally efficient and that it robustly leads to accurate policies.
****

Implicit λ-Jeffreys Autoencoders: Taking the Best of Both Worlds
Aibek Alanov,Max Kochurov,Artem Sobolev,Daniil Yashkov,Dmitry Vetrov
We propose a new form of an autoencoding model which incorporates the best properties of variational autoencoders (VAE) and generative adversarial networks (GAN). It is known that GAN can produce very realistic samples while VAE does not suffer from mode collapsing problem. Our model optimizes λ-Jeffreys divergence between the model distribution and the true data distribution. We show that it takes the best properties of VAE and GAN objectives. It consists of two parts. One of these parts can be optimized by using the standard adversarial training, and the second one is the very objective of the VAE model. However, the straightforward way of substituting the VAE loss does not work well if we use an explicit likelihood such as Gaussian or Laplace which have limited flexibility in high dimensions and are unnatural for modelling images in the space of pixels. To tackle this problem we propose a novel approach to train the VAE model with an implicit likelihood by an adversarially trained discriminator. In an extensive set of experiments on CIFAR-10 and TinyImagent datasets, we show that our model achieves the state-of-the-art generation and reconstruction quality and demonstrate how we can balance between mode-seeking and mode-covering behaviour of our model by adjusting the weight λ in our objective.
****

FLUID FLOW MASS TRANSPORT FOR GENERATIVE NETWORKS
Jingrong Lin,Keegan Lensink,Eldad Haber
Generative Adversarial Networks have been shown to be powerful tools for generating content resulting in them being intensively studied in recent years. Training these networks requires maximizing a generator loss and minimizing a discriminator loss, leading to a difficult saddle point problem that is slow and difficult to converge. Motivated by techniques in the registration of point clouds and the fluid flow formulation of mass transport, we investigate a new formulation that is based on strict minimization, without the need for the maximization. This formulation views the problem as a matching problem rather than an adversarial one, and thus allows us to quickly converge and obtain meaningful metrics in the optimization path.
****

LEX-GAN: Layered Explainable Rumor Detector Based on Generative Adversarial Networks
Mingxi Cheng,Yizhi Li,Shahin Nazarian,Paul Bogdan
Social media have emerged to be increasingly popular and have been used as tools for gathering and propagating information. However, the vigorous growth of social media contributes to the fast-spreading and far-reaching rumors. Rumor detection has become a necessary defense. Traditional rumor detection methods based on hand-crafted feature selection are replaced by automatic approaches that are ba

sed on Artificial Intelligence (AI). AI decision making systems need to have the necessary means, such as explainability to assure users their trustworthiness. Inspired by the thriving development of Generative Adversarial Networks (GANs) on text applications, we propose LEX-GAN, a GAN-based layered explainable rumor detector to improve the detection quality and provide explainability. Unlike fake news detection that needs a previously collected verified news database, LEX-GAN realizes explainable rumor detection based on only tweet-level text. LEX-GAN is trained with generated non-rumor-looking rumors. The generators produce rumors by intelligently inserting controversial information in non-rumors, and force the discriminators to detect detailed glitches and deduce exactly which parts in the sentence are problematic. The layered structures in both generative and discriminative model contributes to the high performance. We show LEX-GAN's mutation detection ability in textural sequences by performing a gene classification and mutation detection task.

****

Towards Stable and Efficient Training of Verifiably Robust Neural Networks

Huan Zhang,Hongge Chen,Chaowei Xiao,Sven Gowal,Robert Stanforth,Bo Li,Duane Boning,Cho-Jui Hsieh

Training neural networks with verifiable robustness guarantees is challenging. Several existing approaches utilize linear relaxation based neural network output bounds under perturbation, but they can slow down training by a factor of hundreds depending on the underlying network architectures. Meanwhile, interval bound propagation (IBP) based training is efficient and significantly outperforms linear relaxation based methods on many tasks, yet it may suffer from stability issues since the bounds are much looser especially at the beginning of training. In this paper, we propose a new certified adversarial training method, CROWN-IBP, by combining the fast IBP bounds in a forward bounding pass and a tight linear relaxation based bound, CROWN, in a backward bounding pass. CROWN-IBP is computationally efficient and consistently outperforms IBP baselines on training verifiably robust neural networks. We conduct large scale experiments on MNIST and CIFAR datasets, and outperform all previous linear relaxation and bound propagation based certified defenses in L_inf robustness.

Notably, we achieve 7.02% verified test error on MNIST at epsilon=0.3, and 66.94% on CIFAR-10 with epsilon=8/255.

****

Learning in Confusion: Batch Active Learning with Noisy Oracle

Gaurav Gupta,Anit Kumar Sahu,Wan-Yi Lin

We study the problem of training machine learning models incrementally using active learning with access to imperfect or noisy oracles. We specifically consider the setting of batch active learning, in which multiple samples are selected as opposed to a single sample as in classical settings so as to reduce the training overhead. Our approach bridges between uniform randomness and score based importance sampling of clusters when selecting a batch of new samples. Experiments on
benchmark image classification datasets (MNIST, SVHN, and CIFAR10) shows improvement over existing active learning strategies. We introduce an extra denoising layer to deep networks to make active learning robust to label noises and show significant improvements.

****

Iterative energy-based projection on a normal data manifold for anomaly localization

David Dehaene,Oriel Frigo,Sébastien Combrexelle,Pierre Eline

Autoencoder reconstructions are widely used for the task of unsupervised anomaly localization. Indeed, an autoencoder trained on normal data is expected to only be able to reconstruct normal features of the data, allowing the segmentation of anomalous pixels in an image via a simple comparison between the image and its autoencoder reconstruction. In practice however, local defects added to a normal image can deteriorate the whole reconstruction, making this segmentation challenging. To tackle the issue, we propose in this paper a new approach for project

ing anomalous data on a autoencoder-learned normal data manifold, by using gradient descent on an energy derived from the autoencoder's loss function. This energy can be augmented with regularization terms that model priors on what constitutes the user-defined optimal projection. By iteratively updating the input of the autoencoder, we bypass the loss of high-frequency information caused by the autoencoder bottleneck. This allows to produce images of higher quality than classic reconstructions. Our method achieves state-of-the-art results on various anomaly localization datasets. It also shows promising results at an inpainting task on the CelebA dataset.
****

Chart Auto-Encoders for Manifold Structured Data
Stephan Schonsheck,Jie Chen,Rongjie Lai
Auto-encoding and generative models have made tremendous successes in image and signal representation learning and generation. These models, however, generally employ the full Euclidean space or a bounded subset (such as $[0,1]^l$) as the latent space, whose trivial geometry is often too simplistic to meaningfully reflect the structure of the data. This paper aims at exploring a nontrivial geometric structure of the latent space for better data representation. Inspired by differential geometry, we propose \textbf{Chart Auto-Encoder (CAE)}, which captures the manifold structure of the data with multiple charts and transition functions among them. CAE translates the mathematical definition of manifold through parameterizing the entire data set as a collection of overlapping charts, creating local latent representations. These representations are an enhancement of the single-charted latent space commonly employed in auto-encoding models, as they reflect the intrinsic structure of the manifold. Therefore, CAE achieves a more accurate approximation of data and generates realistic new ones. We conduct experiments with synthetic and real-life data to demonstrate the effectiveness of the proposed CAE.
****

Optimizing Loss Landscape Connectivity via Neuron Alignment
N. Joseph Tatro,Pin-Yu Chen,Payel Das,Igor Melnyk,Prasanna Sattigeri,Rongjie Lai
The loss landscapes of deep neural networks are poorly understood due to their high nonconvexity. Empirically, the local optima of these loss functions can be connected by a simple curve in model space, along which the loss remains fairly constant. Yet, current path finding algorithms do not consider the influence of symmetry in the loss surface caused by weight permutations of the networks corresponding to the minima. We propose a framework to investigate the effect of symmetry on the landscape connectivity by directly optimizing the weight permutations of the networks being connected. Through utilizing an existing neuron alignment technique, we derive an initialization for the weight permutations. Empirically, this initialization is critical for efficiently learning a simple, planar, low-loss curve between networks that successfully generalizes. Additionally, we introduce a proximal alternating minimization scheme to address if an optimal permutation can be learned, with some provable convergence guarantees. We find that the learned parameterized curve is still a low-loss curve after permuting the weights of the endpoint models, for a subset of permutations. We also show that there is small but steady performance gain in performance of the ensembles constructed from the learned curve, when considering weight space symmetry.
****

CROSS-DOMAIN CASCADED DEEP TRANSLATION
Oren Katzir,Dani Lischinski,Daniel Cohen-Or
In recent years we have witnessed tremendous progress in unpaired image-to-image translation methods, propelled by the emergence of DNNs and adversarial training strategies. However, most existing methods focus on transfer of style and appearance, rather than on shape translation. The latter task is challenging, due to its intricate non-local nature, which calls for additional supervision. We mitigate this by descending the deep layers of a pre-trained network, where the deep features contain more semantics, and applying the translation between these deep features. Specifically, we leverage VGG, which is a classification network, pre-trained with large-scale semantic supervision. Our translation is performed in

a cascaded, deep-to-shallow, fashion, along the deep feature hierarchy: we firs t translate between the deepest layers that encode the higher-level semantic con tent of the image, proceeding to translate the shallower layers, conditioned on the deeper ones. We show that our method is able to translate between different domains, which exhibit significantly different shapes. We evaluate our method bo th qualitatively and quantitatively and compare it to state-of-the-art image-to-image translation methods. Our code and trained models will be made available.
****

V1Net: A computational model of cortical horizontal connections
Vijay Veerabadran,Virginia R. de Sa
The primate visual system builds robust, multi-purpose representations of the ex ternal world in order to support several diverse downstream cortical processes. Such representations are required to be invariant to the sensory inconsistencies caused by dynamically varying lighting, local texture distortion, etc. A key ar chitectural feature combating such environmental irregularities is ‘long-range h orizontal connections’ that aid the perception of the global form of objects. In this work, we explore the introduction of such horizontal connections into stan dard deep convolutional networks; we present V1Net -- a novel convolutional-recu rrent unit that models linear and nonlinear horizontal inhibitory and excitatory connections inspired by primate visual cortical connectivity. We introduce the Texturized Challenge -- a new benchmark to evaluate object recognition performan ce under perceptual noise -- which we use to evaluate V1Net against an array of carefully selected control models with/without recurrent processing. Additionall y, we present results from an ablation study of V1Net demonstrating the utility of diverse neurally inspired horizontal connections for state-of-the-art AI syst ems on the task of object boundary detection from natural images. We also presen t the emergence of several biologically plausible horizontal connectivity patter ns, namely center-on surround-off, association fields and border-ownership conne ctivity patterns in a V1Net model trained to perform boundary detection on natur al images from the Berkeley Segmentation Dataset 500 (BSDS500). Our findings sug gest an increased representational similarity between V1Net and biological visua l systems, and highlight the importance of neurally inspired recurrent contextua l processing principles for learning visual representations that are robust to p erceptual noise and furthering the state-of-the-art in computer vision.
****

Distribution Matching Prototypical Network for Unsupervised Domain Adaptation
Lei Zhu,Wei Wang,Mei Hui Zhang,Beng Chin Ooi,Chang Yao
State-of-the-art Unsupervised Domain Adaptation (UDA) methods learn transferable features by minimizing the feature distribution discrepancy between the source and target domains. Different from these methods which do not model the feature distributions explicitly, in this paper, we explore explicit feature distributio n modeling for UDA. In particular, we propose Distribution Matching Prototypical Network (DMPN) to model the deep features from each domain as Gaussian mixture distributions. With explicit feature distribution modeling, we can easily measur e the discrepancy between the two domains. In DMPN, we propose two new domain di screpancy losses with probabilistic interpretations. The first one minimizes the distances between the corresponding Gaussian component means of the source and target data. The second one minimizes the pseudo negative log likelihood of gene rating the target features from source feature distribution. To learn both discr iminative and domain invariant features, DMPN is trained by minimizing the class ification loss on the labeled source data and the domain discrepancy losses toge ther. Extensive experiments are conducted over two UDA tasks. Our approach yield s a large margin in the Digits Image transfer task over state-of-the-art approac hes. More remarkably, DMPN obtains a mean accuracy of 81.4% on VisDA 2017 datase t. The hyper-parameter sensitivity analysis shows that our approach is robust w. r.t hyper-parameter changes.
****

Deep amortized clustering
Juho Lee,Yoonho Lee,Yee Whye Teh
We propose a \textit{deep amortized clustering} (DAC), a neural architecture whi

ch learns to cluster datasets efficiently using a few forward passes. DAC implic
itly learns what makes a cluster, how to group data points into clusters, and ho
w to count the number of clusters in datasets. DAC is meta-learned using labelle
d datasets for training, a process distinct from traditional clustering algorith
ms which usually require hand-specified prior knowledge about cluster shapes/str
uctures. We empirically show, on both synthetic and image data, that DAC can eff
iciently and accurately cluster new datasets coming from the same distribution u
sed to generate training datasets.
****

Using Objective Bayesian Methods to Determine the Optimal Degree of Curvature wi
thin the Loss Landscape
Devon Jarvis,Richard Klein,Benjamin Rosman
The efficacy of the width of the basin of attraction surrounding a minimum in pa
rameter space as an indicator for the generalizability of a model parametrizatio
n is a point of contention surrounding the training of artificial neural network
s, with the dominant view being that wider areas in the landscape reflect better
 generalizability by the trained model. In this work, however, we aim to show th
at this is only true for a noiseless system and in general the trend of the mode
l towards wide areas in the landscape reflect the propensity of the model to ove
rfit the training data. Utilizing the objective Bayesian (Jeffreys) prior we ins
tead propose a different determinant of the optimal width within the parameter l
andscape determined solely by the curvature of the landscape. In doing so we uti
lize the decomposition of the landscape into the dimensions of principal curvatu
re and find the first principal curvature dimension of the parameter space to be
 independent of noise within the training data.
****

Towards neural networks that provably know when they don't know
Alexander Meinke,Matthias Hein
It has recently been shown that ReLU networks produce arbitrarily over-confident
 predictions far away from the
training data. Thus, ReLU networks do not know when they don't know. However, th
is is a highly important property in safety
critical applications. In the context of out-of-distribution detection (OOD) the
re have been a number of proposals to mitigate this problem but none of them are
 able to make any mathematical guarantees. In this paper we propose a new approa
ch to OOD which overcomes both problems. Our approach can be used with ReLU netw
orks and provides provably low confidence predictions far away from the training
 data as well as the first certificates for low confidence predictions in a neig
hborhood of an out-distribution point. In the experiments we show that state-of-
the-art methods fail in this worst-case setting whereas our model can guarantee
its performance while retaining state-of-the-art OOD performance.
****

BatchEnsemble: an Alternative Approach to Efficient Ensemble and Lifelong Learni
ng
Yeming Wen,Dustin Tran,Jimmy Ba

Ensembles, where multiple neural networks are trained individually and their pre
dictions are averaged, have been shown to be widely successful for improving bot
h the accuracy and predictive uncertainty of single neural networks. However, an
 ensemble’s cost for both training and testing increases linearly with the numbe
r of networks, which quickly becomes untenable.
In this paper, we propose BatchEnsemble, an ensemble method whose computational
and memory costs are significantly lower than typical ensembles. BatchEnsemble a
chieves this by defining each weight matrix to be the Hadamard product of a shar
ed weight among all ensemble members and a rank-one matrix per member. Unlike en
sembles, BatchEnsemble is not only parallelizable across devices, where one devi
ce trains one member, but also parallelizable within a device, where multiple en
semble members are updated simultaneously for a given mini-batch. Across CIFAR-1
0, CIFAR-100, WMT14 EN-DE/EN-FR translation, and out-of-distribution tasks, Batc
hEnsemble yields competitive accuracy and uncertainties as typical ensembles; th

e speedup at test time is 3X and memory reduction is 3X at an ensemble of size 4 . We also apply BatchEnsemble to lifelong learning, where on Split-CIFAR-100, Ba tchEnsemble yields comparable performance to progressive neural networks while h aving a much lower computational and memory costs. We further show that BatchEns emble can easily scale up to lifelong learning on Split-ImageNet which involves 100 sequential learning tasks
****

Fully Convolutional Graph Neural Networks using Bipartite Graph Convolutions
Marcel Nassar,Xin Wang,Evren Tumer
Graph neural networks have been adopted in numerous applications ranging from le arning relational representations to modeling data on irregular domains such as point clouds, social graphs, and molecular structures. Though diverse in nature, graph neural network architectures remain limited by the graph convolution oper ator whose input and output graphs must have the same structure. With this restr iction, representational hierarchy can only be built by graph convolution operat ions followed by non-parameterized pooling or expansion layers. This is very muc h like early convolutional network architectures, which later have been replaced by more effective parameterized strided and transpose convolution operations in combination with skip connections. In order to bring a similar change to graph convolutional networks, here we introduce the bipartite graph convolution operat ion, a parameterized transformation between different input and output graphs. O ur framework is general enough to subsume conventional graph convolution and poo ling as its special cases and supports multi-graph aggregation leading to a clas s of flexible and adaptable network architectures, termed BiGraphNet. By replaci ng the sequence of graph convolution and pooling in hierarchical architectures w ith a single parametric bipartite graph convolution, (i) we answer the question of whether graph pooling matters, and (ii) accelerate computations and lower mem ory requirements in hierarchical networks by eliminating pooling layers. Then, w ith concrete examples, we demonstrate that the general BiGraphNet formalism (iii) provides the modeling flexibility to build efficient architectures such as gra ph skip connections, and autoencoders.
****

Inductive representation learning on temporal graphs
da Xu,chuanwei ruan,evren korpeoglu,sushant kumar,kannan achan
Inductive representation learning on temporal graphs is an important step toward salable machine learning on real-world dynamic networks. The evolving nature of temporal dynamic graphs requires handling new nodes as well as capturing tempor al patterns. The node embeddings, which are now functions of time, should repres ent both the static node features and the evolving topological structures. Moreo ver, node and topological features can be temporal as well, whose patterns the n ode embeddings should also capture. We propose the temporal graph attention (TGA T) layer to efficiently aggregate temporal-topological neighborhood features to learn the time-feature interactions. For TGAT, we use the self-attention mechani sm as building block and develop a novel functional time encoding technique base d on the classical Bochner's theorem from harmonic analysis. By stacking TGAT la yers, the network recognizes the node embeddings as functions of time and is abl e to inductively infer embeddings for both new and observed nodes as the graph e volves. The proposed approach handles both node classification and link predicti on task, and can be naturally extended to include the temporal edge features. We evaluate our method with transductive and inductive tasks under temporal settin gs with two benchmark and one industrial dataset. Our TGAT model compares favora bly to state-of-the-art baselines as well as the previous temporal graph embeddi ng approaches.
****

Attention on Abstract Visual Reasoning
Lukas Hahne,Timo Lüddecke,Florentin Wörgötter,David Kappel
Attention mechanisms have been boosting the performance of deep learning models on a wide range of applications, ranging from speech understanding to program in duction. However, despite experiments from psychology which suggest that attent ion plays an essential role in visual reasoning, the full potential of attention

mechanisms has so far not been explored to solve abstract cognitive tasks on image data. In this work, we propose a hybrid network architecture, grounded on self-attention and relational reasoning. We call this new model Attention Relation Network (ARNe). ARNe combines features from the recently introduced Transformer and the Wild Relation Network (WReN). We test ARNe on the Procedurally Generated Matrices (PGMs) datasets for abstract visual reasoning. ARNe excels the WReN model on this task by 11.28 ppt. Relational concepts between objects are efficiently learned demanding only 35% of the training samples to surpass reported accuracy of the base line model. Our proposed hybrid model, represents an alternative on learning abstract relations using self-attention and demonstrates that the Transformer network is also well suited for abstract visual reasoning.
****
Starfire: Regularization-Free Adversarially-Robust Structured Sparse Training
Noah Gamboa,Kais Kudrolli,Anand Dhoot,Ardavan Pedram
This paper studies structured sparse training of CNNs with a gradual pruning technique that leads to fixed, sparse weight matrices after a set number of epochs. We simplify the structure of the enforced sparsity so that it reduces overhead caused by regularization. The proposed training methodology explores several options for structured sparsity.

We study various tradeoffs with respect to pruning duration, learning-rate configuration, and the total length of training.
We show that our method creates a sparse version of ResNet50 and ResNet50v1.5 on full ImageNet while remaining within a negligible <1% margin of accuracy loss.
To make sure that this type of sparse training does not harm the robustness of the network, we also demonstrate how the network behaves in the presence of adversarial attacks. Our results show that with 70% target sparsity, over 75% top-1 accuracy is achievable.
****
Convolutional Tensor-Train LSTM for Long-Term Video Prediction
Jiahao Su,Wonmin Byeon,Furong Huang,Jan Kautz,Animashree Anandkumar
Long-term video prediction is highly challenging since it entails simultaneously capturing spatial and temporal information across a long range of image frames. Standard recurrent models are ineffective since they are prone to error propagation and cannot effectively capture higher-order correlations. A potential solution is to extend to higher-order spatio-temporal recurrent models. However, such a model requires a large number of parameters and operations, making it intractable to learn in practice and is prone to overfitting. In this work, we propose convolutional tensor-train LSTM (Conv-TT-LSTM), which learns higher-orderConvolutional LSTM (ConvLSTM) efficiently using convolutional tensor-train decomposition (CTTD). Our proposed model naturally incorporates higher-order spatio-temporal information at a small cost of memory and computation by using efficient low-rank tensor representations. We evaluate our model on Moving-MNIST and KTH datasets and show improvements over standard ConvLSTM and better/comparable results to other ConvLSTM-based approaches, but with much fewer parameters.
****
An Information Theoretic Approach to Distributed Representation Learning
Abdellatif Zaidi,Inaki Estella Aguerri
The problem of distributed representation learning is one in which multiple sources of information X1,...,XK are processed separately so as to extract useful information about some statistically correlated ground truth Y. We investigate this problem from information-theoretic grounds. For both discrete memoryless (DM) and memoryless vector Gaussian models, we establish fundamental limits of learning in terms of optimal tradeoffs between accuracy and complexity. We also develop a variational bound on the optimal tradeoff that generalizes the evidence lower bound (ELBO) to the distributed setting. Furthermore, we provide a variational inference type algorithm that allows to compute this bound and in which the mappings are parametrized by neural networks and the bound approximated by Markov sampling and optimized with stochastic gradient descent. Experimental results on synthetic and real datasets are provided to support the efficiency of the approa

ches and algorithms which we develop in this paper.
****
A Probabilistic Formulation of Unsupervised Text Style Transfer
Junxian He,Xinyi Wang,Graham Neubig,Taylor Berg-Kirkpatrick
We present a deep generative model for unsupervised text style transfer that unifies previously proposed non-generative techniques. Our probabilistic approach models non-parallel data from two domains as a partially observed parallel corpus. By hypothesizing a parallel latent sequence that generates each observed sequence, our model learns to transform sequences from one domain to another in a completely unsupervised fashion. In contrast with traditional generative sequence models (e.g. the HMM), our model makes few assumptions about the data it generates: it uses a recurrent language model as a prior and an encoder-decoder as a transduction distribution. While computation of marginal data likelihood is intractable in this model class, we show that amortized variational inference admits a practical surrogate. Further, by drawing connections between our variational objective and other recent unsupervised style transfer and machine translation techniques, we show how our probabilistic view can unify some known non-generative objectives such as backtranslation and adversarial loss. Finally, we demonstrate the effectiveness of our method on a wide range of unsupervised style transfer tasks, including sentiment transfer, formality transfer, word decipherment, author imitation, and related language translation. Across all style transfer tasks, our approach yields substantial gains over state-of-the-art non-generative baselines, including the state-of-the-art unsupervised machine translation techniques that our approach generalizes. Further, we conduct experiments on a standard unsupervised machine translation task and find that our unified approach matches the current state-of-the-art.
****
ROBUST GENERATIVE ADVERSARIAL NETWORK
Shufei Zhang,Zhuang Qian,Kaizhu Huang,Rui Zhang,Jimin Xiao
Generative adversarial networks (GANs) are powerful generative models, but usually suffer from instability which may lead to poor generations. Most existing works try to alleviate this problem by focusing on stabilizing the training of the discriminator, which unfortunately ignores the robustness of generator and discriminator. In this work, we consider the robustness of GANs and propose a novel robust method called robust generative adversarial network (RGAN). Particularly, we design a robust optimization framework where the generator and discriminator compete with each other in a worst-case setting within a small Wasserstein ball. The generator tries to map the worst input distribution (rather than a specific input distribution, typically a Gaussian distribution used in most GANs) to the real data distribution, while the discriminator attempts to distinguish the real and fake distribution with the worst perturbation. We have provided theories showing that the generalization of the new robust framework can be guaranteed. A series of experiments on CIFAR-10, STL-10 and CelebA datasets indicate that our proposed robust framework can improve consistently on four baseline GAN models. We also provide ablation analysis and visualization showing the efficacy of our method on both generator and discriminator quantitatively and qualitatively.
****
Feature Map Transform Coding for Energy-Efficient CNN Inference
Brian Chmiel,Chaim Baskin,Ron Banner,Evgenii Zheltonozhskii,Yevgeny Yermolin,Alex Karbachevsky,Alex M. Bronstein,Avi Mendelson
    Convolutional neural networks (CNNs) achieve state-of-the-art accuracy in a variety of tasks in computer vision and beyond. One of the major obstacles hindering the ubiquitous use of CNNs for inference on low-power edge devices is their high computational complexity and memory bandwidth requirements. The latter often dominates the energy footprint on modern hardware. In this paper, we introduce a lossy transform coding approach, inspired by image and video compression, designed to reduce the memory bandwidth due to the storage of intermediate activation calculation results. Our method does not require fine-tuning the network weights and halves the data transfer volumes to the main memory by compressing feature maps, which are highly correlated, with variable length coding. Our method

outperform previous approach in term of the number of bits per value with minor accuracy degradation on ResNet-34 and MobileNetV2. We analyze the performance of our approach on a variety of CNN architectures and demonstrate that FPGA implementation of ResNet-18 with our approach results in a reduction of around 40% in the memory energy footprint, compared to quantized network, with negligible impact on accuracy. When allowing accuracy degradation of up to 2%, the reduction of 60% is achieved. A reference implementation}accompanies the paper.
****

Generative Models for Effective ML on Private, Decentralized Datasets
Sean Augenstein,H. Brendan McMahan,Daniel Ramage,Swaroop Ramaswamy,Peter Kairouz,Mingqing Chen,Rajiv Mathews,Blaise Aguera y Arcas
To improve real-world applications of machine learning, experienced modelers develop intuition about their datasets, their models, and how the two interact. Manual inspection of raw data—of representative samples, of outliers, of misclassifications—is an essential tool in a) identifying and fixing problems in the data, b) generating new modeling hypotheses,
and c) assigning or refining human-provided labels. However, manual data inspection is risky for privacy-sensitive datasets, such as those representing the behavior of real-world individuals. Furthermore, manual data inspection is impossible in the increasingly important setting of federated learning, where raw examples are stored at the edge and the modeler may only access aggregated outputs such as metrics or model parameters. This paper demonstrates that generative models—trained using federated methods and with formal differential privacy guarantees—can be used effectively to debug data issues even
when the data cannot be directly inspected. We explore these methods in applications to text with differentially private federated RNNs and to images using a novel algorithm for differentially private federated GANs.
****

Learning from Partially-Observed Multimodal Data with Variational Autoencoders
Yu Gong,Hossein Hajimirsadeghi,Jiawei He,Megha Nawhal,Thibaut Durand,Greg Mori
Learning from only partially-observed data for imputation has been an active research area. Despite promising progress on unimodal data imputation (e.g., image in-painting), models designed for multimodal data imputation are far from satisfactory. In this paper, we propose variational selective autoencoders (VSAE) for this task. Different from previous works, our proposed VSAE learns only from partially-observed data. The proposed VSAE is capable of learning the joint distribution of observed and unobserved modalities as well as the imputation mask, resulting in a unified model for various down-stream tasks including data generation and imputation.
Evaluation on both synthetic high-dimensional and challenging low-dimensional multi-modality datasets shows significant improvement over the state-of-the-art data imputation models.
****

A SIMPLE AND EFFECTIVE FRAMEWORK FOR PAIRWISE DEEP METRIC LEARNING
Qi Qi,Yan Yan,Zixuan Wu,Xiaoyu Wang,Tianbao Yang
Deep metric learning (DML) has received much attention in deep learning due to its wide applications in computer vision. Previous studies have focused on designing complicated losses and hard example mining methods, which are mostly heuristic and lack of theoretical understanding. In this paper, we cast DML as a simple pairwise binary classification problem that classifies a pair of examples as similar or dissimilar. It identifies the most critical issue in this problem---imbalanced data pairs. To tackle this issue, we propose a simple and effective framework to sample pairs in a batch of data for updating the model. The key to this framework is to define a robust loss for all pairs over a mini-batch of data, which is formulated by distributionally robust optimization. The flexibility in constructing the {\it uncertainty decision set} of the dual variable allows us to recover state-of-the-art complicated losses and also to induce novel variants.
Empirical studies on several benchmark data sets demonstrate that our simple and effective method outperforms the state-of-the-art results.
****

A Group-Theoretic Framework for Knowledge Graph Embedding
Tong Yang,Long Sha,Pengyu Hong
We have rigorously proved the existence of a group algebraic structure hidden in relational knowledge embedding problems, which suggests that a group-based embedding framework is essential for model design. Our theoretical analysis explores merely the intrinsic property of the embedding problem itself without introducing extra designs. Using the proposed framework, one could construct embedding models that naturally accommodate all possible local graph patterns, which are necessary for reproducing a complete graph from atomic knowledge triplets. We reconstruct many state-of-the-art models from the framework and re-interpret them as embeddings with different groups. Moreover, we also propose new instantiation models using simple continuous non-abelian groups.
****
A■MCTS: SEARCH WITH THEORETICAL GUARANTEE USING POLICY AND VALUE FUNCTIONS
Xian Wu,Yuandong Tian,Lexing Ying
Combined with policy and value neural networks, Monte Carlos Tree Search (MCTS) is a critical component of the recent success of AI agents in learning to play board games like Chess and Go (Silver et al., 2017). However, the theoretical foundations of MCTS with policy and value networks remains open. Inspired by MCTS, we propose A■MCTS, a novel search algorithm that uses both the policy and value predictors to guide search and enjoys theoretical guarantees. Specifically, assuming that value and policy networks give reasonably accurate signals of the values of each state and action, the sample complexity (number of calls to the value network) to estimate the value of the current state, as well as the optimal one-step action to take from the current state, can be bounded. We apply our theoretical framework to different models for the noise distribution of the policy and value network as well as the distribution of rewards, and show that for these general models, the sample complexity is polynomial in D, where D is the depth of the search tree. Empirically, our method outperforms MCTS in these models.
****
Picking Winning Tickets Before Training by Preserving Gradient Flow
Chaoqi Wang,Guodong Zhang,Roger Grosse
Overparameterization has been shown to benefit both the optimization and generalization of neural networks, but large networks are resource hungry at both training and test time. Network pruning can reduce test-time resource requirements, but is typically applied to trained networks and therefore cannot avoid the expensive training process. We aim to prune networks at initialization, thereby saving resources at training time as well. Specifically, we argue that efficient training requires preserving the gradient flow through the network. This leads to a simple but effective pruning criterion we term Gradient Signal Preservation (GraSP). We empirically investigate the effectiveness of the proposed method with extensive experiments on CIFAR-10, CIFAR-100, Tiny-ImageNet and ImageNet, using VGGNet and ResNet architectures. Our method can prune 80% of the weights of a VGG-16 network on ImageNet at initialization, with only a 1.6% drop in top-1 accuracy. Moreover, our method achieves significantly better performance than the baseline at extreme sparsity levels. Our code is made public
at: https://github.com/alecwangcq/GraSP.
****
Exploring Cellular Protein Localization Through Semantic Image Synthesis
Daniel Li,Qiang Ma,Andrew Liu,Justin Cheung,Dana Pe’er,Itsik Pe’er
Cell-cell interactions have an integral role in tumorigenesis as they are critical in governing immune responses. As such, investigating specific cell-cell interactions has the potential to not only expand upon the understanding of tumorigenesis, but also guide clinical management of patient responses to cancer immunotherapies. A recent imaging technique for exploring cell-cell interactions, multiplexed ion beam imaging by time-of-flight (MIBI-TOF), allows for cells to be quantified in 36 different protein markers at sub-cellular resolutions in situ as high resolution multiplexed images. To explore the MIBI images, we propose a GAN for multiplexed data with protein specific attention. By conditioning image generation on cell types, sizes, and neighborhoods through semantic segmentation map

s, we are able to observe how these factors affect cell-cell interactions simult aneously in different protein channels. Furthermore, we design a set of metrics and offer the first insights towards cell spatial orientations, cell protein exp ressions, and cell neighborhoods. Our model, cell-cell interaction GAN (CCIGAN), outperforms or matches existing image synthesis methods on all conventional mea sures and significantly outperforms on biologically motivated metrics. To our kn owledge, we are the first to systematically model multiple cellular protein beha viors and interactions under simulated conditions through image synthesis.

****

Learning Calibratable Policies using Programmatic Style-Consistency

Eric Zhan,Albert Tseng,Yisong Yue,Adith Swaminathan,Matthew Hausknecht

We study the important and challenging problem of controllable generation of lon g-term sequential behaviors. Solutions to this problem would impact many applica tions, such as calibrating behaviors of AI agents in games or predicting player trajectories in sports. In contrast to the well-studied areas of controllable ge neration of images, text, and speech, there are significant challenges that are unique to or exacerbated by generating long-term behaviors: how should we specif y the factors of variation to control, and how can we ensure that the generated temporal behavior faithfully demonstrates diverse styles? In this paper, we leve rage large amounts of raw behavioral data to learn policies that can be calibrat ed to generate a diverse range of behavior styles (e.g., aggressive versus passi ve play in sports). Inspired by recent work on leveraging programmatic labeling functions, we present a novel framework that combines imitation learning with da ta programming to learn style-calibratable policies. Our primary technical contr ibution is a formal notion of style-consistency as a learning objective, and its integration with conventional imitation learning approaches. We evaluate our fr amework using demonstrations from professional basketball players and agents in the MuJoCo physics environment, and show that our learned policies can be accura tely calibrated to generate interesting behavior styles in both domains.

****

Contextual Temperature for Language Modeling

Pei-Hsin Wang,Sheng-Iou Hsieh,Shieh-Chieh Chang,Jia-Yu Pan,Yu-Ting Chen,Wei Wei, Da-Cheng Juan

Temperature scaling has been widely used to improve performance for NLP tasks th at utilize Softmax decision layer. Current practices in using temperature either assume a ■xed value or a dynamically changing temperature but with a ■xed sched ule. Little has been known on an optimal trajectory of temperature that can chan ge with the context. In this paper, we propose contextual temperature, a mechani sm that allows temperatures to change over the context for each vocabulary, and to co-adopt with model parameters during training. Experimental results illustra ted that contextual temperature improves over state-of-the-art language models s igni■cantly. Our model CT-MoS achieved a perplexity of 55.31 in the test set of Penn Treebank and a perplexity of 62.89 in the test set of WikiText-2. The in-de pth analysis showed that the behavior of temperature schedule varies dramaticall y by vocabulary. The optimal temperature trajectory drops as the context becomes longer to suppress uncertainties in language modeling. These evidence further j usti■ed the need for contextual temperature and explained its performance advant age over ■xed temperature or scheduling.

****

Retrospection: Leveraging the Past for Efficient Training of Deep Neural Network s

Ayush Chopra,Surgan Jandial,Mausoom Sarkar,Balaji Krishnamurthy,Vineeth Balasubr amanian

Deep neural networks are powerful learning machines that have enabled breakthrou ghs in several domains. In this work, we introduce retrospection loss to improve the performance of neural networks by utilizing prior experiences during traini ng. Minimizing the retrospection loss pushes the parameter state at the current training step towards the optimal parameter state while pulling it away from the parameter state at a previous training step. We conduct extensive experiments t o show that the proposed retrospection loss results in improved performance acro

ss multiple tasks, input types and network architectures.
****
Curriculum Loss: Robust Learning and Generalization against Label Corruption
Yueming Lyu,Ivor W. Tsang
Deep neural networks (DNNs) have great expressive power, which can even memorize samples with wrong labels. It is vitally important to reiterate robustness and generalization in DNNs against label corruption. To this end, this paper studies the 0-1 loss, which has a monotonic relationship between empirical adversary (reweighted) risk (Hu et al. 2018). Although the 0-1 loss is robust to outliers, it is also difficult to optimize. To efficiently optimize the 0-1 loss while keeping its robust properties, we propose a very simple and efficient loss, i.e. curriculum loss (CL). Our CL is a tighter upper bound of the 0-1 loss compared with conventional summation based surrogate losses. Moreover, CL can adaptively select samples for stagewise training. As a result, our loss can be deemed as a novel perspective of curriculum sample selection strategy, which bridges a connection between curriculum learning and robust learning. Experimental results on noisy MNIST, CIFAR10 and CIFAR100 dataset validate the robustness of the proposed loss.
****
Adversarially Robust Generalization Just Requires More Unlabeled Data
Runtian Zhai,Tianle Cai,Di He,Chen Dan,Kun He,John E. Hopcroft,Liwei Wang
Neural network robustness has recently been highlighted by the existence of adversarial examples. Many previous works show that the learned networks do not perform well on perturbed test data, and significantly more labeled data is required to achieve adversarially robust generalization. In this paper, we theoretically and empirically show that with just more unlabeled data, we can learn a model with better adversarially robust generalization. The key insight of our results is based on a risk decomposition theorem, in which the expected robust risk is separated into two parts: the stability part which measures the prediction stability in the presence of perturbations, and the accuracy part which evaluates the standard classification accuracy. As the stability part does not depend on any label information, we can optimize this part using unlabeled data. We further prove that for a specific Gaussian mixture problem, adversarially robust generalization can be almost as easy as the standard generalization in supervised learning if a sufficiently large amount of unlabeled data is provided. Inspired by the theoretical findings, we further show that a practical adversarial training algorithm that leverages unlabeled data can improve adversarial robust generalization on MNIST and Cifar-10.
****
Why Does the VQA Model Answer No?: Improving Reasoning through Visual and Linguistic Inference
Seungjun Jung,Junyoung Byun,Kyujin Shim,Changick Kim
In order to make Visual Question Answering (VQA) explainable, previous studies not only visualize the attended region of a VQA model but also generate textual explanations for its answers. However, when the model's answer is 'no,' existing methods have difficulty in revealing detailed arguments that lead to that answer. In addition, previous methods are insufficient to provide logical bases, when the question requires common sense to answer. In this paper, we propose a novel textual explanation method to overcome the aforementioned limitations. First, we extract keywords that are essential to infer an answer from a question. Second, for a pre-trained explanation generator, we utilize a novel Variable-Constrained
Beam Search (VCBS) algorithm to generate phrases that best describes the relationship between keywords in images. Then, we complete an explanation by feeding the phrase to the generator. Furthermore, if the answer to the question is "yes" or "no," we apply Natural Langauge Inference (NLI) to identify whether contents of the question can be inferred from the explanation using common sense. Our user study, conducted in Amazon Mechanical Turk (MTurk), shows that our proposed method generates more reliable explanations compared to the previous methods. Moreover, by modifying the VQA model's answer through the output of the NLI model, we

show that VQA performance increases by 1.1% from the original model.
****
DeepSFM: Structure From Motion Via Deep Bundle Adjustment
Xingkui Wei,Yinda Zhang,Zhuwen Li,Yanwei Fu,Xiangyang Xue
Structure from motion (SfM) is an essential computer vision problem which has not been well handled by deep learning. One of the promising trends is to apply explicit structural constraint, e.g. 3D cost volume, into the network. In this work, we design a physical driven architecture, namely DeepSFM, inspired by traditional Bundle Adjustment (BA), which consists of two cost volume based architectures for depth and pose estimation respectively, iteratively running to improve both. In each cost volume, we encode not only photo-metric consistency across multiple input images, but also geometric consistency to ensure that depths from multiple views agree with each other. The explicit constraints on both depth (structure) and pose (motion), when combined with the learning components, bring the merit from both traditional BA and emerging deep learning technology. Extensive experiments on various datasets show that our model achieves the state-of-the-art performance on both depth and pose estimation with superior robustness against less number of inputs and the noise in initialization.

****
IsoNN: Isomorphic Neural Network for Graph Representation Learning and Classification
Lin Meng,Jiawei Zhang
Deep learning models have achieved huge success in numerous fields, such as computer vision and natural language processing. However, unlike such fields, it is hard to apply traditional deep learning models on the graph data due to the ‘node-orderless’ property. Normally, adjacency matrices will cast an artificial and random node-order on the graphs, which renders the performance of deep mod- els on graph classification tasks extremely erratic, and the representations learned by such models lack clear interpretability. To eliminate the unnecessary node-order constraint, we propose a novel model named Isomorphic Neural Network (ISONN), which learns the graph representation by extracting its isomorphic features via the graph matching between input graph and templates. ISONN has two main components: graph isomorphic feature extraction component and classification component. The graph isomorphic feature extraction component utilizes a set of subgraph templates as the kernel variables to learn the possible subgraph patterns existing in the input graph and then computes the isomorphic features. A set of permutation matrices is used in the component to break the node-order brought by the matrix representation. Three fully-connected layers are used as the classification component in ISONN. Extensive experiments are conducted on benchmark datasets, the experimental results can demonstrate the effectiveness of ISONN, especially compared with both classic and state-of-the-art graph classification methods.
****
Uncertainty-guided Continual Learning with Bayesian Neural Networks
Sayna Ebrahimi,Mohamed Elhoseiny,Trevor Darrell,Marcus Rohrbach
Continual learning aims to learn new tasks without forgetting previously learned ones. This is especially challenging when one cannot access data from previous tasks and when the model has a fixed capacity. Current regularization-based continual learning algorithms need an external representation and extra computation to measure the parameters' \textit{importance}. In contrast, we propose Uncertainty-guided Continual Bayesian Neural Networks (UCB), where the learning rate adapts according to the uncertainty defined in the probability distribution of the weights in networks. Uncertainty is a natural way to identify \textit{what to remember} and \textit{what to change} as we continually learn, and thus mitigate catastrophic forgetting. We also show a variant of our model, which uses uncertainty for weight pruning
and retains task performance after pruning by saving binary masks per tasks. We evaluate our UCB approach extensively on diverse object classification datasets with short and long sequences of tasks and report superior or on-par performance compared to existing approaches. Additionally, we show that our model does not

necessarily need task information at test time, i.e. it does not presume knowled
ge of which task a sample belongs to.
****

On Empirical Comparisons of Optimizers for Deep Learning
Dami Choi,Christopher J. Shallue,Zachary Nado,Jaehoon Lee,Chris J. Maddison,Geor
ge E. Dahl
Selecting an optimizer is a central step in the contemporary deep learning pipel
ine. In this paper we demonstrate the sensitivity of optimizer comparisons to th
e metaparameter tuning protocol. Our findings suggest that the metaparameter sea
rch space may be the single most important factor explaining the rankings obtain
ed by recent empirical comparisons in the literature. In fact, we show that thes
e results can be contradicted when metaparameter search spaces are changed. As t
uning effort grows without bound, more general update rules should never underpe
rform the ones they can approximate (i.e., Adam should never perform worse than
momentum), but the recent attempts to compare optimizers either assume these inc
lusion relationships are not relevant in practice or restrict the metaparameters
 they tune to break the inclusions. In our experiments, we find that the inclusi
on relationships between optimizers matter in practice and always predict optimi
zer comparisons. In particular, we find that the popular adative gradient method
s never underperform momentum or gradient descent. We also report practical tips
 around tuning rarely-tuned metaparameters of adaptive gradient methods and rais
e concerns about fairly benchmarking optimizers for neural network training.
****

On Evaluating Explainability Algorithms
Gokula Krishnan Santhanam,Ali Alami-Idrissi,Nuno Mota,Anika Schumann,Ioana Giurg
iu
A plethora of methods attempting to explain predictions of black-box models have
 been proposed by the Explainable Artificial Intelligence (XAI) community. Yet,
measuring the quality of the generated explanations is largely unexplored, makin
g quantitative comparisons non-trivial. In this work, we propose a suite of mult
ifaceted metrics that enables us to objectively compare explainers based on the
correctness, consistency, as well as the confidence of the generated explanation
s. These metrics are computationally inexpensive, do not require model-retrainin
g and can be used across different data modalities. We evaluate them on common e
xplainers such as Grad-CAM, SmoothGrad, LIME and Integrated Gradients. Our exper
iments show that the proposed metrics reflect qualitative observations reported
in earlier works.
****

Deep Hierarchical-Hyperspherical Learning (DH^2L)
Youngsung Kim,Jae-Joon Han
Regularization is known to be an inexpensive and reasonable solution to alleviat
e over-fitting problems of inference models, including deep neural networks. In
this paper, we propose a hierarchical regularization which preserves the semanti
c structure of a sample distribution. At the same time, this regularization prom
otes diversity by imposing distance between parameter vectors enlarged within se
mantic structures. To generate evenly distributed parameters, we constrain them
to lie on \emph{hierarchical hyperspheres}. Evenly distributed parameters are co
nsidered to be less redundant. To define hierarchical parameter space, we propos
e to reformulate the topology space with multiple hypersphere space. On each hyp
ersphere space, the projection parameter is defined by two individual parameters
. Since maximizing groupwise pairwise distance between points on hypersphere is
nontrivial (generalized Thomson problem), we propose a new discrete metric integ
rated with continuous angle metric. Extensive experiments on publicly available
datasets (CIFAR-10, CIFAR-100, CUB200-2011, and Stanford Cars), our proposed met
hod shows improved generalization performance, especially when the number of su
per-classes is larger.
****

Versatile Anomaly Detection with Outlier Preserving Distribution Mapping Autoenc
oders
Walter Gerych,Elke Rundensteiner,Emmanuel Agu

State-of-the-art deep learning methods for outlier detection make the assumption that anomalies will appear far away from inlier data in the latent space produced by distribution mapping deep networks. However, this assumption fails in practice, because the divergence penalty adopted for this purpose encourages mapping outliers into the same high-probability regions as inliers. To overcome this shortcoming, we introduce a novel deep learning outlier detection method, called Outlier Preserving Distribution Mapping Autoencoder (OP-DMA), which succeeds to map outliers to low probability regions in the latent space of an autoencoder. For this we leverage the insight that outliers are likely to have a higher reconstruction error than inliers. We thus achieve outlier-preserving distribution mapping through weighting the reconstruction error of individual points by the value of a multivariate Gaussian probability density function evaluated at those points. This weighting implies that outliers will result overall penalty if they are mapped to low-probability regions. We show that if the global minimum of our newly proposed loss function is achieved,
then our OP-DMA maps inliers to regions with a Mahalanobis distance less than delta, and outliers to regions past this delta, delta being the inverse Chi Squared CDF evaluated at (1-alpha) with alpha the percentage of outliers in the data set. Our experiments confirm that OP-DMA consistently outperforms the state-of-art methods on a rich variety of outlier detection benchmark datasets.
****

Ladder Polynomial Neural Networks
Li-Ping Liu,Ruiyuan Gu,Xiaozhe Hu
The underlying functions of polynomial neural networks are polynomial functions. These networks are shown to have nice theoretical properties by previous analysis, but they are actually hard to train when their polynomial orders are high. In this work, we devise a new type of activations and then create the Ladder Polynomial Neural Network (LPNN). This new network
can be trained with generic optimization algorithms. With a feedforward structure, it can also be combined with deep learning techniques such as batch normalization and dropout. Furthermore, an LPNN provides good control of its polynomial order because its polynomial order increases by 1 with each of its hidden layers. In our empirical study, deep LPNN models achieve good performances in a series of regression and classification tasks.
****

Training Recurrent Neural Networks Online by Learning Explicit State Variables
Somjit Nath,Vincent Liu,Alan Chan,Xin Li,Adam White,Martha White
Recurrent neural networks (RNNs) allow an agent to construct a state-representation from a stream of experience, which is essential in partially observable problems. However, there are two primary issues one must overcome when training an RNN: the sensitivity of the learning algorithm's performance to truncation length and and long training times. There are variety of strategies to improve training in RNNs, the mostly notably Backprop Through Time (BPTT) and by Real-Time Recurrent Learning. These strategies, however, are typically computationally expensive and focus computation on computing gradients back in time. In this work, we reformulate the RNN training objective to explicitly learn state vectors; this breaks the dependence across time and so avoids the need to estimate gradients far back in time. We show that for a fixed buffer of data, our algorithm---called Fixed Point Propagation (FPP)---is sound: it converges to a stationary point of the new objective. We investigate the empirical performance of our online FPP algorithm, particularly in terms of computation compared to truncated BPTT with varying truncation levels.
****

Improved Modeling of Complex Systems Using Hybrid Physics/Machine Learning/Stochastic Models
Anand Ramakrishnan,Warren B. Jackson,Kent Evans
Combining domain knowledge models with neural models has been challenging. End-to-end trained neural models often perform better (lower Mean Square Error) than domain knowledge models or domain/neural combinations, and the combination is inefficient to train. In this paper, we demonstrate that by composing domain mod

els with machine learning models, by using extrapolative testing sets, and invok ing decorrelation objective functions, we create models which can predict more c omplex systems. The models are interpretable, extrapolative, data-efficient, and capture predictable but complex non-stochastic behavior such as unmodeled degre es of freedom and systemic measurement noise. We apply this improved modeling p aradigm to several simulated systems and an actual physical system in the contex t of system identification. Several ways of composing domain models with neura l models are examined for time series, boosting, bagging, and auto-encoding on v arious systems of varying complexity and non-linearity. Although this work is p reliminary, we show that the ability to combine models is a very promising direc tion for neural modeling.

****

LEARNING TO LEARN WITH BETTER CONVERGENCE

Patrick H. Chen,Sashank Reddi,Sanjiv Kumar,Cho-Jui Hsieh

We consider the learning to learn problem, where the goal is to leverage deeplea rning models to automatically learn (iterative) optimization algorithms for training machine learning models. A natural way to tackle this problem is to replace the human-designed optimizer by an LSTM network and train the parameter s on some simple optimization problems (Andrychowicz et al., 2016). Despite the ir success compared to traditional optimizers such as SGD on a short horizon, th eselearnt (meta-) optimizers suffer from two key deficiencies: they fail to conv erge(or can even diverge) on a longer horizon (e.g., 10000 steps). They also oft en fail to generalize to new tasks. To address the convergence problem, we rethi nk the architecture design of the meta-optimizer and develop an embarrassingly s imple,yet powerful form of meta-optimizers—a coordinate-wise RNN model. We provi de insights into the problems with the previous designs of each component and re -design our SimpleOptimizer to resolve those issues. Furthermore, we propose ane w mechanism to allow information sharing between coordinates which enables the m eta-optimizer to exploit second-order information with negligible overhead.With these designs, our proposed SimpleOptimizer outperforms previous meta-optimizers and can successfully converge to optimal solutions in the long run.Furthermore, our empirical results show that these benefits can be obtained with much smalle r models compared to the previous ones.

****

Deep Expectation-Maximization in Hidden Markov Models via Simultaneous Perturbat ion Stochastic Approximation

Chong Li,Dan Shen,C.J. Richard Shi,Hongxia Yang

We propose a novel method to estimate the parameters of a collection of Hidden M arkov Models (HMM), each of which corresponds to a set of known features. The ob servation sequence of an individual HMM is noisy and/or insufficient, making par ameter estimation solely based on its corresponding observation sequence a chall enging problem. The key idea is to combine the classical Expectation-Maximizatio n (EM) algorithm with a neural network, while these two are jointly trained in a n end-to-end fashion, mapping the HMM features to its parameters and effectively fusing the information across different HMMs. In order to address the numerical difficulty in computing the gradient of the EM iteration, simultaneous perturba tion stochastic approximation (SPSA) is employed to approximate the gradient. We also provide a rigorous proof that the approximated gradient due to SPSA conver ges to the true gradient almost surely. The efficacy of the proposed method is d emonstrated on synthetic data as well as a real-world e-Commerce dataset.

****

Cross-lingual Alignment vs Joint Training: A Comparative Study and A Simple Unif ied Framework

Zirui Wang*,Jiateng Xie*,Ruochen Xu,Yiming Yang,Graham Neubig,Jaime G. Carbonell

Learning multilingual representations of text has proven a successful method for many cross-lingual transfer learning tasks. There are two main paradigms for le arning such representations: (1) alignment, which maps different independently t rained monolingual representations into a shared space, and (2) joint training, which directly learns unified multilingual representations using monolingual and cross-lingual objectives jointly. In this paper, we first conduct direct compar

isons of representations learned using both of these methods across diverse cros s-lingual tasks. Our empirical results reveal a set of pros and cons for both me thods, and show that the relative performance of alignment versus joint training is task-dependent. Stemming from this analysis, we propose a simple and novel f ramework that combines these two previously mutually-exclusive approaches. Exten sive experiments demonstrate that our proposed framework alleviates limitations of both approaches, and outperforms existing methods on the MUSE bilingual lexic on induction (BLI) benchmark. We further show that this framework can generalize to contextualized representations such as Multilingual BERT, and produces state -of-the-art results on the CoNLL cross-lingual NER benchmark.
****

Compositional Visual Generation with Energy Based Models
Yilun Du,Shuang Li,Igor Mordatch
Humans are able to both learn quickly and rapidly adapt their knowledge. One maj or component is the ability to incrementally combine many simple concepts to acc elerates the learning process. We show that energy based models are a promising class of models towards exhibiting these properties by directly combining probab ility distributions. This allows us to combine an arbitrary number of different distributions in a globally coherent manner. We show this compositionality prope rty allows us to define three basic operators, logical conjunction, disjunction, and negation, on different concepts to generate plausible naturalistic images. Furthermore, by applying these abilities, we show that we are able to extrapolat e concept combinations, continually combine previously learned concepts, and inf er concept properties in a compositional manner.
****

Hierarchical Bayes Autoencoders
Shuangfei Zhai,Carlos Guestrin,Joshua M. Susskind
Autoencoders are powerful generative models for complex data, such as images. Ho wever, standard models like the variational autoencoder (VAE) typically have uni modal Gaussian decoders, which cannot effectively represent the possible semanti c variations in the space of images. To address this problem, we present a new p robabilistic generative model called the \emph{Hierarchical Bayes Autoencoder (H BAE)}. The HBAE contains a multimodal decoder in the form of an energy-based mod el (EBM), instead of the commonly adopted unimodal Gaussian distribution. The HB AE can be trained using variational inference, similar to a VAE, to recover late nt codes conditioned on inputs. For the decoder, we use an adversarial approxima tion where a conditional generator is trained to match the EBM distribution. Dur ing inference time, the HBAE consists of two sampling steps: first a latent code for the input is sampled, and then this code is passed to the conditional gener ator to output a stochastic reconstruction. The HBAE is also capable of modeling sets, by inferring a latent code for a set of examples, and sampling set member s through the multimodal decoder. In both single image and set cases, the decode r generates plausible variations consistent with the input data, and generates r ealistic unconditional samples. To the best our knowledge, Set-HBAE is the first model that is able to generate complex image sets.
****

Wyner VAE: A Variational Autoencoder with Succinct Common Representation Learnin g
J. Jon Ryu,Yoojin Choi,Young-Han Kim,Mostafa El-Khamy,Jungwon Lee
A new variational autoencoder (VAE) model is proposed that learns a succinct com mon representation of two correlated data variables for conditional and joint ge neration tasks. The proposed Wyner VAE model is based on two information theoret ic problems---distributed simulation and channel synthesis---in which Wyner's co mmon information arises as the fundamental limit of the succinctness of the comm on representation. The Wyner VAE decomposes a pair of correlated data variables into their common representation (e.g., a shared concept) and local representat ions that capture the remaining randomness (e.g., texture and style) in respecti ve data variables by imposing the mutual information between the data variables and the common representation as a regularization term. The utility of the propo sed approach is demonstrated through experiments for joint and conditional gener

ation with and without style control using synthetic data and real images. Experimental results show that learning a succinct common representation achieves better generative performance and that the proposed model outperforms existing VAE variants and the variational information bottleneck method.
****
Granger Causal Structure Reconstruction from Heterogeneous Multivariate Time Series
Yunfei Chu,Xiaowei Wang,Chunyan Feng,Jianxin Ma,Jingren Zhou,Hongxia Yang
Granger causal structure reconstruction is an emerging topic that can uncover causal relationship behind multivariate time series data. In many real-world systems, it is common to encounter a large amount of multivariate time series data collected from heterogeneous individuals with sharing commonalities, however there are ongoing concerns regarding its applicability in such large scale complex scenarios, presenting both challenges and opportunities for Granger causal reconstruction. To bridge this gap, we propose a Granger cAusal StructurE Reconstruction (GASER) framework for inductive Granger causality learning and common causal structure detection on heterogeneous multivariate time series. In particular, we address the problem through a novel attention mechanism, called prototypical Granger causal attention. Extensive experiments, as well as an online A/B test on an E-commercial advertising platform, demonstrate the superior performances of GASER.
****
CGT: Clustered Graph Transformer for Urban Spatio-temporal Prediction
Xu Geng,Lingyu Zhang,Shulin Li,Yuanbo Zhang,Lulu Zhang,Leye Wang,Qiang Yang,Hongtu Zhu,Jieping Ye
Deep learning based approaches have been widely used in various urban spatio-temporal forecasting problems, but most of them fail to account for the unsmoothness issue of urban data in their architecture design, which significantly deteriorates their prediction performance. The aim of this paper is to develop a novel clustered graph transformer framework that integrates both graph attention network and transformer under an encoder-decoder architecture to address such unsmoothness issue. Specifically, we propose two novel structural components to refine the architectures of those existing deep learning models. In spatial domain, we propose a gradient-based clustering method to distribute different feature extractors to regions in different contexts. In temporal domain, we propose to use multi-view position encoding to address the periodicity and closeness of urban time series data. Experiments on real datasets obtained from a ride-hailing business show that our method can achieve 10\%-25\% improvement than many state-of-the-art baselines.
****
Robust Reinforcement Learning for Continuous Control with Model Misspecification
Daniel J. Mankowitz,Nir Levine,Rae Jeong,Abbas Abdolmaleki,Jost Tobias Springenberg,Yuanyuan Shi,Jackie Kay,Todd Hester,Timothy Mann,Martin Riedmiller
We provide a framework for incorporating robustness -- to perturbations in the transition dynamics which we refer to as model misspecification -- into continuous control Reinforcement Learning (RL) algorithms. We specifically focus on incorporating robustness into a state-of-the-art continuous control RL algorithm called Maximum a-posteriori Policy Optimization (MPO). We achieve this by learning a policy that optimizes for a worst case, entropy-regularized, expected return objective and derive a corresponding robust entropy-regularized Bellman contraction operator. In addition, we introduce a less conservative, soft-robust, entropy-regularized objective with a corresponding Bellman operator. We show that both, robust and soft-robust policies, outperform their non-robust counterparts in nine Mujoco domains with environment perturbations. In addition, we show improved robust performance on a challenging, simulated, dexterous robotic hand. Finally, we present multiple investigative experiments that provide a deeper insight into the robustness framework; including an adaptation to another continuous control RL algorithm. Performance videos can be found online at https://sites.google.com/view/robust-rl.
****

Decoupling Representation and Classifier for Long-Tailed Recognition
Bingyi Kang,Saining Xie,Marcus Rohrbach,Zhicheng Yan,Albert Gordo,Jiashi Feng,Yannis Kalantidis
The long-tail distribution of the visual world poses great challenges for deep learning based classification models on how to handle the class imbalance problem. Existing solutions usually involve class-balancing strategies, e.g., by loss re-weighting, data re-sampling, or transfer learning from head- to tail-classes, but most of them adhere to the scheme of jointly learning representations and classifiers. In this work, we decouple the learning procedure into representation learning and classification, and systematically explore how different balancing strategies affect them for long-tailed recognition. The findings are surprising: (1) data imbalance might not be an issue in learning high-quality representations; (2) with representations learned with the simplest instance-balanced (natural) sampling, it is also possible to achieve strong long-tailed recognition ability by adjusting only the classifier. We conduct extensive experiments and set new state-of-the-art performance on common long-tailed benchmarks like ImageNet-LT, Places-LT and iNaturalist, showing that it is possible to outperform carefully designed losses, sampling strategies, even complex modules with memory, by using a straightforward approach that decouples representation and classification. Our code is available at https://github.com/facebookresearch/classifier-balancing.
****
SDGM: Sparse Bayesian Classifier Based on a Discriminative Gaussian Mixture Model
Hideaki Hayashi,Seiichi Uchida
In probabilistic classification, a discriminative model based on Gaussian mixture exhibits flexible fitting capability. Nevertheless, it is difficult to determine the number of components. We propose a sparse classifier based on a discriminative Gaussian mixture model (GMM), which is named sparse discriminative Gaussian mixture (SDGM). In the SDGM, a GMM-based discriminative model is trained by sparse Bayesian learning. This learning algorithm improves the generalization capability by obtaining a sparse solution and automatically determines the number of components by removing redundant components. The SDGM can be embedded into neural networks (NNs) such as convolutional NNs and can be trained in an end-to-end manner. Experimental results indicated that the proposed method prevented overfitting by obtaining sparsity. Furthermore, we demonstrated that the proposed method outperformed a fully connected layer with the softmax function in certain cases when it was used as the last layer of a deep NN.
****
Which Tasks Should Be Learned Together in Multi-task Learning?
Trevor Standley,Amir R. Zamir,Dawn Chen,Leonidas Guibas,Jitendra Malik,Silvio Savarese
Many computer vision applications require solving multiple tasks in real-time. A neural network can be trained to solve multiple tasks simultaneously using 'multi-task learning'. This saves computation at inference time as only a single network needs to be evaluated. Unfortunately, this often leads to inferior overall performance as task objectives compete, which consequently poses the question: which tasks should and should not be learned together in one network when employing multi-task learning? We systematically study task cooperation and competition and propose a framework for assigning tasks to a few neural networks such that cooperating tasks are computed by the same neural network, while competing tasks are computed by different networks. Our framework offers a time-accuracy trade-off and can produce better accuracy using less inference time than not only a single large multi-task neural network but also many single-task networks.

****
COMBINED FLEXIBLE ACTIVATION FUNCTIONS FOR DEEP NEURAL NETWORKS
Renlong Jie,Junbin Gao,Andrey Vasnev,Minh-Ngoc Tran
Activation in deep neural networks is fundamental to achieving non-linear mappings. Traditional studies mainly focus on finding fixed activations for a particul

ar set of learning tasks or model architectures. The research on flexible activation is quite limited in both designing philosophy and application scenarios. In this study, we propose a general combined form of flexible activation functions as well as three principles of choosing flexible activation component. Based on this, we develop two novel flexible activation functions that can be implemented in LSTM cells and auto-encoder layers. Also two new regularisation terms based on assumptions as prior knowledge are proposed. We find that LSTM and auto-encoder models with proposed flexible activations provides significant improvements on time series forecasting and image compressing tasks, while layer-wise regularization can improve the performance of CNN (LeNet-5) models with RPeLu activation in image classification tasks.

****

Teacher-Student Compression with Generative Adversarial Networks

Ruishan Liu,Nicolo Fusi,Lester Mackey

More accurate machine learning models often demand more computation and memory at test time, making them difficult to deploy on CPU- or memory-constrained devices. Teacher-student compression (TSC), also known as distillation, alleviates this burden by training a less expensive student model to mimic the expensive teacher model while maintaining most of the original accuracy. However, when fresh data is unavailable for the compression task, the teacher's training data is typically reused, leading to suboptimal compression. In this work, we propose to augment the compression dataset with synthetic data from a generative adversarial network (GAN) designed to approximate the training data distribution. Our GAN-assisted TSC (GAN-TSC) significantly improves student accuracy for expensive models such as large random forests and deep neural networks on both tabular and image datasets. Building on these results, we propose a comprehensive metric—the TSC Score—to evaluate the quality of synthetic datasets based on their induced TSC performance. The TSC Score captures both data diversity and class affinity, and we illustrate its benefits over the popular Inception Score in the context of image classification.

****

Visual Hide and Seek

Boyuan Chen,Shuran Song,Hod Lipson,Carl Vondrick

We train embodied agents to play Visual Hide and Seek where a prey must navigate in a simulated environment in order to avoid capture from a predator. We place a variety of obstacles in the environment for the prey to hide behind, and we only give the agents partial observations of their environment using an egocentric perspective. Although we train the model to play this game from scratch without any prior knowledge of its visual world, experiments and visualizations show that a representation of other agents automatically emerges in the learned representation. Furthermore, we quantitatively analyze how agent weaknesses, such as slower speed, effect the learned policy. Our results suggest that, although agent weaknesses make the learning problem more challenging, they also cause useful features to emerge in the representation.

****

Unsupervised Temperature Scaling: Robust Post-processing Calibration for Domain Shift

Azadeh Sadat Mozafari,Hugo Siqueira Gomes,Christian Gagne

The uncertainty estimation is critical in real-world decision making applications, especially when distributional shift between the training and test data are prevalent. Many calibration methods in the literature have been proposed to improve the predictive uncertainty of DNNs which are generally not well-calibrated. However, none of them is specifically designed to work properly under domain shift condition. In this paper, we propose Unsupervised Temperature Scaling (UTS) as a robust calibration method to domain shift. It exploits test samples to adjust the uncertainty prediction of deep models towards the test distribution. UTS utilizes a novel loss function, weighted NLL, that allows unsupervised calibration. We evaluate UTS on a wide range of model-datasets which shows the possibility of calibration without labels and demonstrate the robustness of UTS compared to other methods (e.g., TS, MC-dropout, SVI, ensembles) in shifted domains.

****
Pareto Optimality in No-Harm Fairness
Natalia Martinez,Martin Bertran,Guillermo Sapiro
Common fairness definitions in machine learning focus on balancing various notions of disparity and utility. In this work we study fairness in the context of risk disparity among sub-populations. We introduce the framework of Pareto-optimal fairness, where the goal of reducing risk disparity gaps is secondary only to the principle of not doing unnecessary harm, a concept that is especially applicable to high-stakes domains such as healthcare. We provide analysis and methodology to obtain maximally-fair no-harm classifiers on finite datasets. We argue that even in domains where fairness at cost is required, no-harm fairness can prove to be the optimal first step. This same methodology can also be applied to any unbalanced classification task, where we want to dynamically equalize the misclassification risks across outcomes without degrading overall performance any more than strictly necessary. We test the proposed methodology on real case-studies of predicting income, ICU patient mortality, classifying skin lesions from images, and assessing credit risk, demonstrating how the proposed framework compares favorably to other traditional approaches.
****
Learning from Unlabelled Videos Using Contrastive Predictive Neural 3D Mapping
Adam W. Harley,Shrinidhi K. Lakshmikanth,Fangyu Li,Xian Zhou,Hsiao-Yu Fish Tung,Katerina Fragkiadaki
Predictive coding theories suggest that the brain learns by predicting observations at various levels of abstraction. One of the most basic prediction tasks is view prediction: how would a given scene look from an alternative viewpoint? Humans excel at this task. Our ability to imagine and fill in missing information is tightly coupled with perception: we feel as if we see the world in 3 dimensions, while in fact, information from only the front surface of the world hits our retinas. This paper explores the role of view prediction in the development of 3D visual recognition. We propose neural 3D mapping networks, which take as input 2.5D (color and depth) video streams captured by a moving camera, and lift them to stable 3D feature maps of the scene, by disentangling the scene content from the motion of the camera. The model also projects its 3D feature maps to novel viewpoints, to predict and match against target views. We propose contrastive prediction losses to replace the standard color regression loss, and show that this leads to better performance on complex photorealistic data. We show that the proposed model learns visual representations useful for (1) semi-supervised learning of 3D object detectors, and (2) unsupervised learning of 3D moving object detectors, by estimating the motion of the inferred 3D feature maps in videos of dynamic scenes. To the best of our knowledge, this is the first work that empirically shows view prediction to be a scalable self-supervised task beneficial to 3D object detection.
****
Dream to Control: Learning Behaviors by Latent Imagination
Danijar Hafner,Timothy Lillicrap,Jimmy Ba,Mohammad Norouzi
Learned world models summarize an agent's experience to facilitate learning complex behaviors. While learning world models from high-dimensional sensory inputs is becoming feasible through deep learning, there are many potential ways for deriving behaviors from them. We present Dreamer, a reinforcement learning agent that solves long-horizon tasks from images purely by latent imagination. We efficiently learn behaviors by propagating analytic gradients of learned state values back through trajectories imagined in the compact state space of a learned world model. On 20 challenging visual control tasks, Dreamer exceeds existing approaches in data-efficiency, computation time, and final performance.
****
From Inference to Generation: End-to-end Fully Self-supervised Generation of Human Face from Speech
Hyeong-Seok Choi,Changdae Park,Kyogu Lee
This work seeks the possibility of generating the human face from voice solely based on the audio-visual data without any human-labeled annotations. To this end

, we propose a multi-modal learning framework that links the inference stage and generation stage. First, the inference networks are trained to match the speaker identity between the two different modalities. Then the pre-trained inference networks cooperate with the generation network by giving conditional information about the voice. The proposed method exploits the recent development of GANs techniques and generates the human face directly from the speech waveform making our system fully end-to-end. We analyze the extent to which the network can naturally disentangle two latent factors that contribute to the generation of a face image one that comes directly from a speech signal and the other that is not related to it and explore whether the network can learn to generate natural human face image distribution by modeling these factors. Experimental results show that the proposed network can not only match the relationship between the human face and speech, but can also generate the high-quality human face sample conditioned on its speech. Finally, the correlation between the generated face and the corresponding speech is quantitatively measured to analyze the relationship between the two modalities.
****

Active Learning Graph Neural Networks via Node Feature Propagation
Yuexin Wu,Yichong Xu,Aarti Singh,Artur Dubrawski,Yiming Yang
Graph Neural Networks (GNNs) for prediction tasks like node classification or edge prediction have received increasing attention in recent machine learning from graphically structured data. However, a large quantity of labeled graphs is difficult to obtain, which significantly limit the true success of GNNs. Although active learning has been widely studied for addressing label-sparse issues with other data types like text, images, etc., how to make it effective over graphs is an open question for research. In this paper, we present the investigation on active learning with GNNs for node classification tasks. Specifically, we propose a new method, which uses node feature propagation followed by K-Medoids clustering of the nodes for instance selection in active learning. With a theoretical bound analysis we justify the design choice of our approach. In our experiments on four benchmark dataset, the proposed method outperforms other representative baseline methods consistently and significantly.
****

Real or Not Real, that is the Question
Yuanbo Xiangli*,Yubin Deng*,Bo Dai*,Chen Change Loy,Dahua Lin
While generative adversarial networks (GAN) have been widely adopted in various topics, in this paper we generalize the standard GAN to a new perspective by treating realness as a random variable that can be estimated from multiple angles. In this generalized framework, referred to as RealnessGAN, the discriminator outputs a distribution as the measure of realness. While RealnessGAN shares similar theoretical guarantees with the standard GAN, it provides more insights on adversarial learning. More importantly, compared to multiple baselines, RealnessGAN provides stronger guidance for the generator, achieving improvements on both synthetic and real-world datasets. Moreover, it enables the basic DCGAN architecture to generate realistic images at 1024*1024 resolution when trained from scratch.
****

Deep Reinforcement Learning with Implicit Human Feedback
Duo Xu,Mohit Agarwal,Raghupathy Sivakumar,Faramarz Fekri
We consider the following central question in the field of Deep Reinforcement Learning (DRL): How can we use implicit human feedback to accelerate and optimize the training of a DRL algorithm? State-of-the-art methods rely on any human feedback to be provided explicitly, requiring the active participation of humans (e.g., expert labeling, demonstrations, etc.). In this work, we investigate an alternative paradigm, where non-expert humans are silently observing (and assessing) the agent interacting with the environment. The human's intrinsic reactions to the agent's behavior is sensed as implicit feedback by placing electrodes on the human scalp and monitoring what are known as event-related electric potentials. The implicit feedback is then used to augment the agent's learning in the RL tasks. We develop a system to obtain and accurately decode the implicit human feed

back (specifically error-related event potentials) for state-action pairs in an Atari-type environment. As a baseline contribution, we demonstrate the feasibility of capturing error-potentials of a human observer watching an agent learning to play several different Atari-games using an electroencephalogram (EEG) cap, and then decoding the signals appropriately and using them as an auxiliary reward function to a DRL algorithm with the intent of accelerating its learning of the game. Building atop the baseline, we then make the following novel contributions in our work:
(i) We argue that the definition of error-potentials is generalizable across different environments; specifically we show that error-potentials of an observer can be learned for a specific game, and the definition used as-is for another game without requiring re-learning of the error-potentials.
(ii) We propose two different frameworks to combine recent advances in DRL into the error-potential based feedback system in a sample-efficient manner, allowing humans to provide implicit feedback while training in the loop, or prior to the training of the RL agent.
(iii) Finally, we scale the implicit human feedback (via ErrP) based RL to reasonably complex environments (games) and demonstrate the significance of our approach through synthetic and real user experiments.

****
Multi-Sample Dropout for Accelerated Training and Better Generalization
Hiroshi Inoue
Dropout is a simple but efficient regularization technique for achieving better generalization of deep neural networks (DNNs); hence it is widely used in tasks based on DNNs. During training, dropout randomly discards a portion of the neurons to avoid overfitting. This paper presents an enhanced dropout technique, which we call multi-sample dropout, for both accelerating training and improving generalization over the original dropout. The original dropout creates a randomly selected subset (called a dropout sample) from the input in each training iteration while the multi-sample dropout creates multiple dropout samples. The loss is calculated for each sample, and then the sample losses are averaged to obtain the final loss. This technique can be easily implemented without implementing a new operator by duplicating a part of the network after the dropout layer while sharing the weights among the duplicated fully connected layers. Experimental results showed that multi-sample dropout significantly accelerates training by reducing the number of iterations until convergence for image classification tasks using the ImageNet, CIFAR-10, CIFAR-100, and SVHN datasets. Multi-sample dropout does not significantly increase computation cost per iteration for deep convolutional networks because most of the computation time is consumed in the convolution layers before the dropout layer, which are not duplicated. Experiments also showed that networks trained using multi-sample dropout achieved lower error rates and losses for both the training set and validation set.

****
MelNet: A Generative Model for Audio in the Frequency Domain
Sean Vasquez,Mike Lewis
Capturing high-level structure in audio waveforms is challenging because a single second of audio spans tens of thousands of timesteps. While long-range dependencies are difficult to model directly in the time domain, we show that they can be more tractably modelled in two-dimensional time-frequency representations such as spectrograms. By leveraging this representational advantage, in conjunction with a highly expressive probabilistic model and a multiscale generation procedure, we design a model capable of generating high-fidelity audio samples which capture structure at timescales which time-domain models have yet to achieve. We demonstrate that our model captures longer-range dependencies than time-domain models such as WaveNet across a diverse set of unconditional generation tasks, including single-speaker speech generation, multi-speaker speech generation, and music generation.
****

Cross Domain Imitation Learning
Kun Ho Kim,Yihong Gu,Jiaming Song,Shengjia Zhao,Stefano Ermon
We study the question of how to imitate tasks across domains with discrepancies such as embodiment and viewpoint mismatch. Many prior works require paired, aligned demonstrations and an additional RL procedure for the task. However, paired, aligned demonstrations are seldom obtainable and RL procedures are expensive. In this work, we formalize the Cross Domain Imitation Learning (CDIL) problem, which encompasses imitation learning in the presence of viewpoint and embodiment mismatch. Informally, CDIL is the process of learning how to perform a task optimally, given demonstrations of the task in a distinct domain. We propose a two step approach to CDIL: alignment followed by adaptation. In the alignment step we execute a novel unsupervised MDP alignment algorithm, Generative Adversarial MDP Alignment (GAMA), to learn state and action correspondences from unpaired, unaligned demonstrations. In the adaptation step we leverage the correspondences to zero-shot imitate tasks across domains. To describe when CDIL is feasible via alignment and adaptation, we introduce a theory of MDP alignability. We experimentally evaluate GAMA against baselines in both embodiment and viewpoint mismatch scenarios where aligned demonstrations don’t exist and show the effectiveness of our approach.
****
Blending Diverse Physical Priors with Neural Networks
Yunhao Ba,Guangyuan Zhao,Achuta Kadambi
Rethinking physics in the era of deep learning is an increasingly important topic. This topic is special because, in addition to data, one can leverage a vast library of physical prior models (e.g. kinematics, fluid flow, etc) to perform more robust inference. The nascent sub-field of physics-based learning (PBL) studies this problem of blending neural networks with physical priors. While previous PBL algorithms have been applied successfully to specific tasks, it is hard to generalize existing PBL methods to a wide range of physics-based problems. Such generalization would require an architecture that can adapt to variations in the correctness of the physics, or in the quality of training data. No such architecture exists. In this paper, we aim to generalize PBL, by making a first attempt to bring neural architecture search (NAS) to the realm of PBL. We introduce a new method known as physics-based neural architecture search (PhysicsNAS) that is a top-performer across a diverse range of quality in the physical model and the dataset.
****
A closer look at network resolution for efficient network design
Taojiannan Yang,Sijie Zhu,Yan Shen,Mi Zhang,Andrew Willis,Chen Chen
There is growing interest in designing lightweight neural networks for mobile and embedded vision applications. Previous works typically reduce computations from the structure level. For example, group convolution based methods reduce computations by factorizing a vanilla convolution into depth-wise and point-wise convolutions. Pruning based methods prune redundant connections in the network structure. In this paper, we explore the importance of network input for achieving optimal accuracy-efficiency trade-off. Reducing input scale is a simple yet effective way to reduce computational cost. It does not require careful network module design, specific hardware optimization and network retraining after pruning. Moreover, different input scales contain different representations to learn. We propose a framework to mutually learn from different input resolutions and network widths. With the shared knowledge, our framework is able to find better width-resolution balance and capture multi-scale representations. It achieves consistently better ImageNet top-1 accuracy over US-Net under different computation constraints, and outperforms the best compound scale model of EfficientNet by 1.5%. The superiority of our framework is also validated on COCO object detection and instance segmentation as well as transfer learning.
****
Efficient Systolic Array Based on Decomposable MAC for Quantized Deep Neural Networks
Ning-Chi Huang,Huan-Jan Chou,Kai-Chiang Wu

Deep Neural Networks (DNNs) have achieved high accuracy in various machine learn ing applications in recent years. As the recognition accuracy of deep learning a pplications increases, reducing the complexity of these neural networks and perf orming the DNN computation on embedded systems or mobile devices become an emerg ing and crucial challenge. Quantization has been presented to reduce the utiliza tion of computational resources by compressing the input data and weights from f loating-point numbers to integers with shorter bit-width. For practical power re duction, it is necessary to operate these DNNs with quantized parameters on appr opriate hardware. Therefore, systolic arrays are adopted to be the major computa tion units for matrix multiplication in DNN accelerators. To obtain a better tra deoff between the precision/accuracy and power consumption, using parameters wit h various bit-widths among different layers within a DNN is an advanced quantiza tion method. In this paper, we propose a novel decomposition strategy to constru ct a low-power decomposable multiplier-accumulator (MAC) for the energy efficien cy of quantized DNNs. In the experiments, when 65% multiplication operations of VGG-16 are operated in shorter bit-width with at most 1% accuracy loss on the CI FAR-10 dataset, our decomposable MAC has 50% energy reduction compared with a no n-decomposable MAC.
****
Improved Image Augmentation for Convolutional Neural Networks by Copyout and Cop yPairing
Philip May
Image augmentation is a widely used technique to improve the performance of conv olutional neural networks (CNNs). In common image shifting, cropping, flipping, shearing and rotating are used for augmentation. But there are more advanced tec hniques like Cutout and SamplePairing.

In this work we present two improvements of the state-of-the-art Cutout and Samp lePairing techniques. Our new method called Copyout takes a square patch of anot her random training image and copies it onto a random location of each image use d for training. The second technique we discovered is called CopyPairing. It com bines Copyout and SamplePairing for further augmentation and even better perform ance.

We apply different experiments with these augmentation techniques on the CIFAR-1 0 dataset to evaluate and compare them under different configurations. In our ex periments we show that Copyout reduces the test error rate by 8.18% compared wit h Cutout and 4.27% compared with SamplePairing. CopyPairing reduces the test err or rate by 11.97% compared with Cutout and 8.21% compared with SamplePairing.

Copyout and CopyPairing implementations are available at https://github.com/anon ym/anonym.
****
On the Evaluation of Conditional GANs
Terrance DeVries,Adriana Romero,Luis Pineda,Graham W. Taylor,Michal Drozdzal
Conditional Generative Adversarial Networks (cGANs) are finding increasingly wid espread use in many application domains. Despite outstanding progress, quantitat ive evaluation of such models often involves multiple distinct metrics to assess different desirable properties, such as image quality, conditional consistency, and intra-conditioning diversity. In this setting, model benchmarking becomes a challenge, as each metric may indicate a different "best" model. In this paper, we propose the Frechet Joint Distance (FJD), which is defined as the Frechet di stance between joint distributions of images and conditioning, allowing it to im plicitly capture the aforementioned properties in a single metric. We conduct pr oof-of-concept experiments on a controllable synthetic dataset, which consistent ly highlight the benefits of FJD when compared to currently established metrics. Moreover, we use the newly introduced metric to compare existing cGAN-based mod els for a variety of conditioning modalities (e.g. class labels, object masks, b ounding boxes, images, and text captions). We show that FJD can be used as a pro mising single metric for model benchmarking.

****
JAUNE: Justified And Unified Neural language Evaluation
Hassan Kané,Yusuf Kocyigit,Ali Abdalla,Pelkins Ajanoh,Mohamed Coulibali
We review the limitations of BLEU and ROUGE -- the most popular metrics used to assess reference summaries against hypothesis summaries, and introduce JAUNE: a set of criteria for what a good metric should behave like and propose concrete ways to use recent Transformers-based Language Models to assess reference summaries against hypothesis summaries.

****
Statistical Adaptive Stochastic Optimization
Pengchuan Zhang,Hunter Lang,Qiang Liu,Lin Xiao
We investigate statistical methods for automatically scheduling the learning rate (step size) in stochastic optimization. First, we consider a broad family of stochastic optimization methods with constant hyperparameters (including the learning rate and various forms of momentum) and derive a general necessary condition for the resulting dynamics to be stationary. Based on this condition, we develop a simple online statistical test to detect (non-)stationarity and use it to automatically drop the learning rate by a constant factor whenever stationarity is detected. Unlike in prior work, our stationarity condition and our statistical test applies to different algorithms without modification. Finally, we propose a smoothed stochastic line-search method that can be used to warm up the optimization process before the statistical test can be applied effectively. This removes the expensive trial and error for setting a good initial learning rate. The combined method is highly autonomous and it attains state-of-the-art training and testing performance in our experiments on several deep learning tasks.
****
Scalable Neural Learning for Verifiable Consistency with Temporal Specifications
Sumanth Dathathri,Johannes Welbl,Krishnamurthy (Dj) Dvijotham,Ramana Kumar,Aditya Kanade,Jonathan Uesato,Sven Gowal,Po-Sen Huang,Pushmeet Kohli
Formal verification of machine learning models has attracted attention recently, and significant progress has been made on proving simple properties like robustness to small perturbations of the input features. In this context, it has also been observed that folding the verification procedure into training makes it easier to train verifiably robust models. In this paper, we extend the applicability of verified training by extending it to (1) recurrent neural network architectures and (2) complex specifications that go beyond simple adversarial robustness, particularly specifications that capture temporal properties like requiring that a robot periodically visits a charging station or that a language model always produces sentences of bounded length. Experiments show that while models trained using standard training often violate desired specifications, our verified training method produces models that both perform well (in terms of test error or reward) and can be shown to be provably consistent with specifications.
****
Model Comparison of Beer data classification using an electronic nose
Mohammed Abdi,Aminat Adebiyi,Andrea Fasoli,Alberto Mannari,Ronald Labby,Luisa Bozano
Olfaction has been and still is an area which is challenging to the research community. Like other senses of the body, there has been a push to replicate the sense of smell to aid in identifying odorous compounds in the form of an electronic nose. At IBM, our team (Cogniscent) has designed a modular sensor board platform based on the artificial olfaction concept we called EVA (Electronic Volatile Analyzer). EVA is an IoT electronic nose device that aims to reproduce olfaction in living begins by integrating an array of partially specific and uniquely selective smell recognition sensors which are directly exposed to the target chemical analyte or the environment. We are exploring a new technique called temperature-controlled oscillation, which gives us virtual array of sensors to represent our signals/ fingerprint. In our study, we run experiments on identifying different types of beers using EVA. In order to successfully carry this classification

task, the entire process starting from preparation of samples, having a consistent protocol of data collection in place all the way to providing the data to be analyzed and input to a machine learning model is very important. On this paper, we will discuss the process of sniffing volatile organic compounds from liquid beer samples and successfully classifying different brands of beers as a pilot test. We researched on different machine learning models in order to get the best classification accuracy for our Beer samples. The best classification accuracy is achieved by using a multi-level perceptron (MLP) artificial neural network (ANN) model, classification of three different brands of beers after splitting one-week data to a training and testing set yielded an accuracy of 97.334. While using separate weeks of data for training and testing set the model yielded an accuracy of 67.812, this is because of drift playing a role in the overall classification process. Using Random forest, the classification accuracy achieved by the model is 0.923. And Decision Tree achieved 0.911.
****

Non-linear System Identification from Partial Observations via Iterative Smoothing and Learning
Kunal Menda,Jean de Becdelièvre,Jayesh K Gupta,Ilan Kroo,Mykel J. Kochenderfer,Zachary Manchester
System identification is the process of building a mathematical model of an unknown system from measurements of its inputs and outputs. It is a key step for model-based control, estimator design, and output prediction. This work presents an algorithm for non-linear offline system identification from partial observations, i.e. situations in which the system's full-state is not directly observable. The algorithm presented, called SISL, iteratively infers the system's full state through non-linear optimization and then updates the model parameters. We test our algorithm on a simulated system of coupled Lorenz attractors, showing our algorithm's ability to identify high-dimensional systems that prove intractable for particle-based approaches. We also use SISL to identify the dynamics of an aerobatic helicopter. By augmenting the state with unobserved fluid states, we learn a model that predicts the acceleration of the helicopter better than state-of-the-art approaches.
****

Evaluating Lossy Compression Rates of Deep Generative Models
Sicong Huang,Alireza Makhzani,Yanshuai Cao,Roger Grosse
Deep generative models have achieved remarkable progress in recent years. Despite this progress, quantitative evaluation and comparison of generative models remains as one of the important challenges. One of the most popular metrics for evaluating generative models is the log-likelihood. While the direct computation of log-likelihood can be intractable, it has been recently shown that the log-likelihood of some of the most interesting generative models such as variational autoencoders (VAE) or generative adversarial networks (GAN) can be efficiently estimated using annealed importance sampling (AIS). In this work, we argue that the log-likelihood metric by itself cannot represent all the different performance characteristics of generative models, and propose to use rate distortion curves to evaluate and compare deep generative models. We show that we can approximate the entire rate distortion curve using one single run of AIS for roughly the same computational cost as a single log-likelihood estimate. We evaluate lossy compression rates of different deep generative models such as VAEs, GANs (and its variants) and adversarial autoencoders (AAE) on MNIST and CIFAR10, and arrive at a number of insights not obtainable from log-likelihoods alone.
****

LambdaNet: Probabilistic Type Inference using Graph Neural Networks
Jiayi Wei,Maruth Goyal,Greg Durrett,Isil Dillig
As gradual typing becomes increasingly popular in languages like Python and TypeScript, there is a growing need to infer type annotations automatically. While type annotations help with tasks like code completion and static error catching, these annotations cannot be fully inferred by compilers and are tedious to annotate by hand. This paper proposes a probabilistic type inference scheme for TypeScript based on a graph neural network. Our approach first uses lightweight sourc

e code analysis to generate a program abstraction called a type dependency graph , which links type variables with logical constraints as well as name and usage information. Given this program abstraction, we then use a graph neural network to propagate information between related type variables and eventually make type predictions. Our neural architecture can predict both standard types, like numb er or string, as well as user-defined types that have not been encountered durin g training. Our experimental results show that our approach outperforms prior wo rk in this space by 14% (absolute) on library types, while having the ability to make type predictions that are out of scope for existing techniques.
****

Variational Autoencoders with Normalizing Flow Decoders
Rogan Morrow,Wei-Chen Chiu
Recently proposed normalizing flow models such as Glow (Kingma & Dhariwal, 2018) have been shown to be able to generate high quality, high dimensional images wi th relatively fast sampling speed. Due to the inherently restrictive design of a rchitecture , however, it is necessary that their model are excessively deep in order to achieve effective training. In this paper we propose to combine Glow mo del with an underlying variational autoencoder in order to counteract this issue . We demonstrate that such our proposed model is competitive with Glow in terms of image quality while requiring far less time for training. Additionally, our m odel achieves state-of-the-art FID score on CIFAR-10 for a likelihood-based mode l.
****

Model-Augmented Actor-Critic: Backpropagating through Paths
Ignasi Clavera,Yao Fu,Pieter Abbeel
Current model-based reinforcement learning approaches use the model simply as a learned black-box simulator to augment the data for policy optimization or value function learning. In this paper, we show how to make more effective use of the model by exploiting its differentiability. We construct a policy optimization a lgorithm that uses the pathwise derivative of the learned model and policy acros s future timesteps. Instabilities of learning across many timesteps are prevente d by using a terminal value function, learning the policy in an actor-critic fas hion. Furthermore, we present a derivation on the monotonic improvement of our o bjective in terms of the gradient error in the model and value function. We show that our approach (i) is consistently more sample efficient than existing state -of-the-art model-based algorithms, (ii) matches the asymptotic performance of m odel-free algorithms, and (iii) scales to long horizons, a regime where typicall y past model-based approaches have struggled.
****

Metagross: Meta Gated Recursive Controller Units for Sequence Modeling
Yi Tay,Yikang Shen,Alvin Chan,Yew Soon Ong
This paper proposes Metagross (Meta Gated Recursive Controller), a new neural se quence modeling unit. Our proposed unit is characterized by recursive parameteri zation of its gating functions, i.e., gating mechanisms of Metagross are control led by instances of itself, which are repeatedly called in a recursive fashion. This can be interpreted as a form of meta-gating and recursively parameterizing a recurrent model. We postulate that our proposed inductive bias provides modeli ng benefits pertaining to learning with inherently hierarchically-structured seq uence data (e.g., language, logical or music tasks). To this end, we conduct ext ensive experiments on recursive logic tasks (sorting, tree traversal, logical in ference), sequential pixel-by-pixel classification, semantic parsing, code gener ation, machine translation and polyphonic music modeling, demonstrating the wide spread utility of the proposed approach, i.e., achieving state-of-the-art (or cl ose) performance on all tasks.
****

Neural Symbolic Reader: Scalable Integration of Distributed and Symbolic Represe ntations for Reading Comprehension
Xinyun Chen,Chen Liang,Adams Wei Yu,Denny Zhou,Dawn Song,Quoc V. Le
Integrating distributed representations with symbolic operations is essential fo r reading comprehension requiring complex reasoning, such as counting, sorting a

nd arithmetics, but most existing approaches are hard to scale to more domains o r more complex reasoning. In this work, we propose the Neural Symbolic Reader (N eRd), which includes a reader, e.g., BERT, to encode the passage and question, a nd a programmer, e.g., LSTM, to generate a program that is executed to produce t he answer. Compared to previous works, NeRd is more scalable in two aspects: (1) domain-agnostic, i.e., the same neural architecture works for different domains ; (2) compositional, i.e., when needed, complex programs can be generated by rec ursively applying the predefined operators, which become executable and interpre table representations for more complex reasoning. Furthermore, to overcome the c hallenge of training NeRd with weak supervision, we apply data augmentation tech niques and hard Expectation-Maximization (EM) with thresholding. On DROP, a chal lenging reading comprehension dataset that requires discrete reasoning, NeRd ach ieves 1.37%/1.18% absolute improvement over the state-of-the-art on EM/F1 metric s. With the same architecture, NeRd significantly outperforms the baselines on M athQA, a math problem benchmark that requires multiple steps of reasoning, by 25 .5% absolute increment on accuracy when trained on all the annotated programs. M ore importantly, NeRd still beats the baselines even when only 20% of the progra m annotations are given.

****

Variational Autoencoders for Highly Multivariate Spatial Point Processes Intensi ties

Baichuan Yuan,Xiaowei Wang,Jianxin Ma,Chang Zhou,Andrea L. Bertozzi,Hongxia Yang

Multivariate spatial point process models can describe heterotopic data over spa ce. However, highly multivariate intensities are computationally challenging due to the curse of dimensionality. To bridge this gap, we introduce a declustering based hidden variable model that leads to an efficient inference procedure via a variational autoencoder (VAE). We also prove that this model is a generalizati on of the VAE-based model for collaborative filtering. This leads to an interest ing application of spatial point process models to recommender systems. Experime ntal results show the method's utility on both synthetic data and real-world dat a sets.

****

Stochastic Mirror Descent on Overparameterized Nonlinear Models

Navid Azizan,Sahin Lale,Babak Hassibi

Most modern learning problems are highly overparameterized, meaning that the mod el has many more parameters than the number of training data points, and as a re sult, the training loss may have infinitely many global minima (in fact, a manif old of parameter vectors that perfectly interpolates the training data). Therefo re, it is important to understand which interpolating solutions we converge to, how they depend on the initialization point and the learning algorithm, and whet her they lead to different generalization performances. In this paper, we study these questions for the family of stochastic mirror descent (SMD) algorithms, of which the popular stochastic gradient descent (SGD) is a special case. Recently it has been shown that, for overparameterized linear models, SMD converges to t he global minimum that is closest (in terms of the Bregman divergence of the mir ror used) to the initialization point, a phenomenon referred to as implicit regu larization. Our contributions in this paper are both theoretical and experimenta l. On the theory side, we show that in the overparameterized nonlinear setting, if the initialization is close enough to the manifold of global optima, SMD with sufficiently small step size converges to a global minimum that is approximatel y the closest global minimum in Bregman divergence, thus attaining approximate i mplicit regularization. For highly overparametrized models, this closeness comes for free: the manifold of global optima is so high dimensional that with high p robability an arbitrarily chosen initialization will be close to the manifold. O n the experimental side, our extensive experiments on the MNIST and CIFAR-10 dat asets, using various initializations, various mirror descents, and various Bregm an divergences, consistently confirms that this phenomenon indeed happens in dee p learning: SMD converges to the closest global optimum to the initialization po int in the Bregman divergence of the mirror used. Our experiments further indica

te that there is a clear difference in the generalization performance of the sol
utions obtained from different SMD algorithms. Experimenting on the CIFAR-10 dat
aset with different regularizers, l1 to encourage sparsity, l2 (yielding SGD) to
 encourage small Euclidean norm, and l10 to discourage large components in the p
arameter vector, consistently and definitively shows that, for small initializat
ion vectors, l10-SMD has better generalization performance than SGD, which in tu
rn has better generalization performance than l1-SMD. This surprising, and perha
ps counter-intuitive, result strongly suggests the importance of a comprehensive
 study of the role of regularization, and the choice of the best regularizer, to
 improve the generalization performance of deep networks.
****

Denoising and Regularization via Exploiting the Structural Bias of Convolutional
 Generators
Reinhard Heckel and Mahdi Soltanolkotabi
Convolutional Neural Networks (CNNs) have emerged as highly successful tools for
 image generation, recovery, and restoration. A major contributing factor to thi
s success is that convolutional networks impose strong prior assumptions about n
atural images. A surprising experiment that highlights this architectural bias t
owards natural images is that one can remove noise and corruptions from a natura
l image without using any training data, by simply fitting (via gradient descent
) a randomly initialized, over-parameterized convolutional generator to the corr
upted image. While this over-parameterized network can fit the corrupted image p
erfectly, surprisingly after a few iterations of gradient descent it generates a
n almost uncorrupted image. This intriguing phenomenon enables state-of-the-art
CNN-based denoising and regularization of other inverse problems. In this paper,
 we attribute this effect to a particular architectural choice of convolutional
networks, namely convolutions with fixed interpolating filters. We then formally
 characterize the dynamics of fitting a two-layer convolutional generator to a n
oisy signal and prove that early-stopped gradient descent denoises/regularizes.
Our proof relies on showing that convolutional generators fit the structured par
t of an image significantly faster than the corrupted portion.
****

Frequency Analysis for Graph Convolution Network
Hoang NT,Takanori Maehara
In this work, we develop quantitative results to the learnablity of a two-layers
 Graph Convolutional Network (GCN). Instead of analyzing GCN under some classes
of functions, our approach provides a quantitative gap between a two-layers GCN
and a two-layers MLP model. Our analysis is based on the graph signal processing
 (GSP) approach, which can provide much more useful insights than the message-pa
ssing computational model. Interestingly, based on our analysis, we have been ab
le to empirically demonstrate a few case when GCN and other state-of-the-art mod
els cannot learn even when true vertex features are extremely low-dimensional. T
o demonstrate our theoretical findings and propose a solution to the aforementio
ned adversarial cases, we build a proof of concept graph neural network model wi
th stacked filters named Graph Filters Neural Network (gfNN).

****

Network Deconvolution
Chengxi Ye,Matthew Evanusa,Hua He,Anton Mitrokhin,Tom Goldstein,James A. Yorke,C
ornelia Fermuller,Yiannis Aloimonos
Convolution is a central operation in Convolutional Neural Networks (CNNs), whic
h applies a kernel to overlapping regions shifted across the image. However, bec
ause of the strong correlations in real-world image data, convolutional kernels
 are in effect re-learning redundant data. In this work, we show that this redun
dancy has made neural network training challenging, and propose network deconvol
ution, a procedure which optimally removes pixel-wise and channel-wise correlati
ons before the data is fed into each layer. Network deconvolution can be efficie
ntly calculated at a fraction of the computational cost of a convolution layer.
We also show that the deconvolution filters in the first layer of the network re
semble the center-surround structure found in biological neurons in the visual r

egions of the brain. Filtering with such kernels results in a sparse representat ion, a desired property that has been missing in the training of neural networks . Learning from the sparse representation promotes faster convergence and superi or results without the use of batch normalization. We apply our network deconvol ution operation to 10 modern neural network models by replacing batch normalizat ion within each. Extensive experiments show that the network deconvolution opera tion is able to deliver performance improvement in all cases on the CIFAR-10, CI FAR-100, MNIST, Fashion-MNIST, Cityscapes, and ImageNet datasets.
****
Revisiting Self-Training for Neural Sequence Generation
Junxian He,Jiatao Gu,Jiajun Shen,Marc'Aurelio Ranzato
Self-training is one of the earliest and simplest semi-supervised methods. The k ey idea is to augment the original labeled dataset with unlabeled data paired wi th the model's prediction (i.e. the pseudo-parallel data). While self-training h as been extensively studied on classification problems, in complex sequence gene ration tasks (e.g. machine translation) it is still unclear how self-training wo rks due to the compositionality of the target space. In this work, we first empi rically show that self-training is able to decently improve the supervised basel ine on neural sequence generation tasks. Through careful examination of the perf ormance gains, we find that the perturbation on the hidden states (i.e. dropout) is critical for self-training to benefit from the pseudo-parallel data, which a cts as a regularizer and forces the model to yield close predictions for similar unlabeled inputs. Such effect helps the model correct some incorrect prediction s on unlabeled data. To further encourage this mechanism, we propose to inject n oise to the input space, resulting in a noisy version of self-training. Empirica l study on standard machine translation and text summarization benchmarks shows that noisy self-training is able to effectively utilize unlabeled data and impro ve the performance of the supervised baseline by a large margin.
****
Generative Cleaning Networks with Quantized Nonlinear Transform for Deep Neura l Network Defense
Jianhe Yuan,Zhihai He
Effective defense of deep neural networks against adversarial attacks remains a challenging problem, especially under white-box attacks.
In this paper, we develop a new generative cleaning network with quantized nonl inear transform for effective defense of deep neural networks. The generative cleaning network, equipped with a trainable quantized nonlinear transform block , is able to destroy the sophisticated noise pattern of adversarial attacks and recover the original image content. The generative cleaning network and attack d etector network are jointly trained using adversarial learning to minimize both perceptual loss and adversarial loss. Our extensive experimental results demons trate that our approach outperforms the state-of-art methods by large margins in both white-box and black-box attacks. For example, it improves the classificati on accuracy for white-box attacks upon the second best method by more than 40\% on the SVHN dataset and more than 20\% on the challenging CIFAR-10 dataset.
****
Mutual Exclusivity as a Challenge for Deep Neural Networks
Kanishk Gandhi,Brenden Lake
Strong inductive biases allow children to learn in fast and adaptable ways. Chil dren use the mutual exclusivity (ME) bias to help disambiguate how words map to referents, assuming that if an object has one label then it does not need anothe r. In this paper, we investigate whether or not standard neural architectures ha ve a ME bias, demonstrating that they lack this learning assumption. Moreover, w e show that their inductive biases are poorly matched to lifelong learning formu lations of classification and translation. We demonstrate that there is a compel ling case for designing neural networks that reason by mutual exclusivity, which remains an open challenge.
****
Meta-Q-Learning
Rasool Fakoor,Pratik Chaudhari,Stefano Soatto,Alexander J. Smola

This paper introduces Meta-Q-Learning (MQL), a new off-policy algorithm for meta
-Reinforcement Learning (meta-RL). MQL builds upon three simple ideas. First, we
 show that Q-learning is competitive with state-of-the-art meta-RL algorithms if
 given access to a context variable that is a representation of the past traject
ory. Second, a multi-task objective to maximize the average reward across the tr
aining tasks is an effective method to meta-train RL policies. Third, past data
from the meta-training replay buffer can be recycled to adapt the policy on a ne
w task using off-policy updates. MQL draws upon ideas in propensity estimation t
o do so and thereby amplifies the amount of available data for adaptation. Exper
iments on standard continuous-control benchmarks suggest that MQL compares favor
ably with the state of the art in meta-RL.
****

CURSOR-BASED ADAPTIVE QUANTIZATION FOR DEEP NEURAL NETWORK
Bapu Li(*),Yanwen Fan(*),Zhiyu Cheng,Yingze Bao (* means equal contribution)
Deep neural network (DNN) has rapidly found many applications in different scena
rios.
However, its large computational cost and memory consumption are barriers
to computing restrained applications. DNN model quantization is a widely
used method to reduce the DNN storage and computation burden by decreasing
the bit width. In this paper, we propose a novel cursor based adaptive quantizat
ion
method using differentiable architecture search (DAS). The multiple bits'
quantization mechanism is formulated as a DAS process with a continuous cursor
that represents the possible quantization bit. The cursor-based DAS adaptively
searches for the desired quantization bit for each layer. The DAS process can
be solved via an alternative approximate optimization process, which is designed
for mixed quantization scheme of a DNN model. We further devise a new loss
function in the search process to simultaneously optimize accuracy and parameter
size of the model. In the quantization step, based on a new strategy, the closes
t
two integers to the cursor are adopted as the bits to quantize the DNN together
to
reduce the quantization noise and avoid the local convergence problem. Comprehen
sive
experiments on benchmark datasets show that our cursor based adaptive
quantization approach achieves the new state-of-the-art for multiple bits' quant
ization
and can efficiently obtain lower size model with comparable or even better
classification accuracy.
****

Natural Image Manipulation for Autoregressive Models Using Fisher Scores
Wilson Yan,Jonathan Ho,Pieter Abbeel
Deep autoregressive models are one of the most powerful models that exist today
which achieve state-of-the-art bits per dim. However, they lie at a strict disad
vantage when it comes to controlled sample generation compared to latent variabl
e models. Latent variable models such as VAEs and normalizing flows allow meanin
gful semantic manipulations in latent space, which autoregressive models do not
have. In this paper, we propose using Fisher scores as a method to extract embed
dings from an autoregressive model to use for interpolation and show that our me
thod provides more meaningful sample manipulation compared to alternate embeddin
gs such as network activations.
****

Towards a Deep Network Architecture for Structured Smoothness
Haroun Habeeb,Oluwasanmi Koyejo
We propose the Fixed Grouping Layer (FGL); a novel feedforward layer designed to
 incorporate the inductive bias of structured smoothness into a deep learning mo
del. FGL achieves this goal by connecting nodes across layers based on spatial s
imilarity. The use of structured smoothness, as implemented by FGL, is motivated
 by applications to structured spatial data, which is, in turn, motivated by dom
ain knowledge. The proposed model architecture outperforms conventional neural n

etwork architectures across a variety of simulated and real datasets with struct
ured smoothness.
****
On the Global Convergence of Training Deep Linear ResNets
Difan Zou,Philip M. Long,Quanquan Gu
We study the convergence of gradient descent (GD) and stochastic gradient descen
t (SGD) for training L-hidden-layer linear residual networks (ResNets). We pro
ve that for training deep residual networks with certain linear transformations
 at input and output layers, which are fixed throughout training, both GD and SGD
 with zero initialization on all hidden weights can converge to the global minim
um of the training loss. Moreover, when specializing to appropriate Gaussian ran
dom linear transformations, GD and SGD provably optimize wide enough deep linear
 ResNets. Compared with the global convergence result of GD for training standar
d deep linear networks \citep{du2019width}, our condition on the neural network
 width is sharper by a factor of $O(\kappa L)$, where κ denotes the condit
ion number of the covariance matrix of the training data. We further propose a m
odified identity input and output transformations, and show that a $(d+k)$-wide
 neural network is sufficient to guarantee the global convergence of GD/SGD, wher
e d,k are the input and output dimensions respectively.
****
A Closer Look at the Optimization Landscapes of Generative Adversarial Networks
Hugo Berard,Gauthier Gidel,Amjad Almahairi,Pascal Vincent,Simon Lacoste-Julien
Generative adversarial networks have been very successful in generative modeling
, however they remain relatively challenging to train compared to standard deep
 neural networks. In this paper, we propose new visualization techniques for the
 optimization landscapes of GANs that enable us to study the game vector field re
sulting from the concatenation of the gradient of both players. Using these vi
sualization techniques we try to bridge the gap between theory and practice by s
howing empirically that the training of GANs exhibits significant rotations arou
nd LSSP, similar to the one predicted by theory on toy examples. Moreover, we pr
ovide empirical evidence that GAN training seems to converge to a stable station
ary point which is a saddle point for the generator loss, not a minimum, while s
till achieving excellent performance.
****
Perceptual Generative Autoencoders
Zijun Zhang,Ruixiang Zhang,Zongpeng Li,Yoshua Bengio,Liam Paull
Modern generative models are usually designed to match target distributions dire
ctly in the data space, where the intrinsic dimensionality of data can be much l
ower than the ambient dimensionality. We argue that this discrepancy may contrib
ute to the difficulties in training generative models. We therefore propose to m
ap both the generated and target distributions to the latent space using the enc
oder of a standard autoencoder, and train the generator (or decoder) to match th
e target distribution in the latent space. The resulting method, perceptual gene
rative autoencoder (PGA), is then incorporated with a maximum likelihood or vari
ational autoencoder (VAE) objective to train the generative model. With maximum
 likelihood, PGAs generalize the idea of reversible generative models to unrestri
cted neural network architectures and arbitrary latent dimensionalities. When co
mbined with VAEs, PGAs can generate sharper samples than vanilla VAEs. Compared
 to other autoencoder-based generative models using simple priors, PGAs achieve s
tate-of-the-art FID scores on CIFAR-10 and CelebA.
****
Simplified Action Decoder for Deep Multi-Agent Reinforcement Learning
Hengyuan Hu,Jakob N Foerster
In recent years we have seen fast progress on a number of benchmark problems in
 AI, with modern methods achieving near or super human performance in Go, Poker a
nd Dota. One common aspect of all of these challenges is that they are by design
 adversarial or, technically speaking, zero-sum. In contrast to these settings,
 success in the real world commonly requires humans to collaborate and communicat
e with others, in settings that are, at least partially, cooperative. In the las
t year, the card game Hanabi has been established as a new benchmark environment

for AI to fill this gap. In particular, Hanabi is interesting to humans since it is entirely focused on theory of mind, i.e. the ability to effectively reason over the intentions, beliefs and point of view of other agents when observing their actions. Learning to be informative when observed by others is an interesting challenge for Reinforcement Learning (RL): Fundamentally, RL requires agents to explore in order to discover good policies. However, when done naively, this randomness will inherently make their actions less informative to others during training. We present a new deep multi-agent RL method, the Simplified Action Decoder (SAD), which resolves this contradiction exploiting the centralized training phase. During training SAD allows other agents to not only observe the (exploratory) action chosen, but agents instead also observe the greedy action of their team mates. By combining this simple intuition with an auxiliary task for state prediction and best practices for multi-agent learning, SAD establishes a new state of the art for 2-5 players on the self-play part of the Hanabi challenge.
****

JAX MD: End-to-End Differentiable, Hardware Accelerated, Molecular Dynamics in Pure Python

Samuel S. Schoenholz,Ekin D. Cubuk

A large fraction of computational science involves simulating the dynamics of particles that interact via pairwise or many-body interactions. These simulations, called Molecular Dynamics (MD), span a vast range of subjects from physics and materials science to biochemistry and drug discovery. Most MD software involves significant use of handwritten derivatives and code reuse across C++, FORTRAN, and CUDA. This is reminiscent of the state of machine learning before automatic differentiation became popular. In this work we bring the substantial advances in software that have taken place in machine learning to MD with JAX, M.D. (JAX MD). JAX MD is an end-to-end differentiable MD package written entirely in Python that can be just-in-time compiled to CPU, GPU, or TPU. JAX MD allows researchers to iterate extremely quickly and lets researchers easily incorporate machine learning models into their workflows. Finally, since all of the simulation code is written in Python, researchers can have unprecedented flexibility in setting up experiments without having to edit any low-level C++ or CUDA code. In addition to making existing workloads easier, JAX MD allows researchers to take derivatives through whole-simulations as well as seamlessly incorporate neural networks into simulations. This paper explores the architecture of JAX MD and its capabilities through several vignettes. Code is available at github.com/jaxmd/jax-md along with an interactive Colab notebook.
****

Biologically inspired sleep algorithm for increased generalization and adversarial robustness in deep neural networks

Timothy Tadros,Giri Krishnan,Ramyaa Ramyaa,Maxim Bazhenov

Current artificial neural networks (ANNs) can perform and excel at a variety of tasks ranging from image classification to spam detection through training on large datasets of labeled data. While the trained network may perform well on similar testing data, inputs that differ even slightly from the training data may trigger unpredictable behavior. Due to this limitation, it is possible to design inputs with very small perturbations that can result in misclassification. These adversarial attacks present a security risk to deployed ANNs and indicate a divergence between how ANNs and humans perform classification. Humans are robust at behaving in the presence of noise and are capable of correctly classifying objects that are noisy, blurred, or otherwise distorted. It has been hypothesized that sleep promotes generalization of knowledge and improves robustness against noise in animals and humans. In this work, we utilize a biologically inspired sleep phase in ANNs and demonstrate the benefit of sleep on defending against adversarial attacks as well as in increasing ANN classification robustness. We compare the sleep algorithm's performance on various robustness tasks with two previously proposed adversarial defenses - defensive distillation and fine-tuning. We report an increase in robustness after sleep phase to adversarial attacks as well as to general image distortions for three datasets: MNIST, CUB200, and a toy dataset. Overall, these results demonstrate the potential for biologically inspired

solutions to solve existing problems in ANNs and guide the development of more robust, human-like ANNs.
****
Distributed Bandit Learning: Near-Optimal Regret with Efficient Communication
Yuanhao Wang,Jiachen Hu,Xiaoyu Chen,Liwei Wang
We study the problem of regret minimization for distributed bandits learning, in which M agents work collaboratively to minimize their total regret under the coordination of a central server. Our goal is to design communication protocols with near-optimal regret and little communication cost, which is measured by the total amount of transmitted data. For distributed multi-armed bandits, we propose a protocol with near-optimal regret and only $O(M\log(MK))$ communication cost, where K is the number of arms. The communication cost is independent of the time horizon T, has only logarithmic dependence on the number of arms, and matches the lower bound except for a logarithmic factor. For distributed d-dimensional linear bandits, we propose a protocol that achieves near-optimal regret and has communication cost of order $O\left(\left(Md+d\log \log d\right)\log T\right)$, which has only logarithmic dependence on T.
****
Is a Good Representation Sufficient for Sample Efficient Reinforcement Learning?
Simon S. Du,Sham M. Kakade,Ruosong Wang,Lin F. Yang
Modern deep learning methods provide effective means to learn good representations. However, is a good representation itself sufficient for sample efficient reinforcement learning? This question has largely been studied only with respect to (worst-case) approximation error, in the more classical approximate dynamic programming literature. With regards to the statistical viewpoint, this question is largely unexplored, and the extant body of literature mainly focuses on conditions which \emph{permit} sample efficient reinforcement learning with little understanding of what are \emph{necessary} conditions for efficient reinforcement learning.
This work shows that, from the statistical viewpoint, the situation is far subtler than suggested by the more traditional approximation viewpoint, where the requirements on the representation that suffice for sample efficient RL are even more stringent. Our main results provide sharp thresholds for reinforcement learning methods, showing that there are hard limitations on what constitutes good function approximation (in terms of the dimensionality of the representation), where we focus on natural representational conditions relevant to value-based, model-based, and policy-based learning. These lower bounds highlight that having a good (value-based, model-based, or policy-based) representation in and of itself is insufficient for efficient reinforcement learning, unless the quality of this approximation passes certain hard thresholds. Furthermore, our lower bounds also imply exponential separations on the sample complexity between 1) value-based learning with perfect representation and value-based learning with a good-but-not-perfect representation, 2) value-based learning and policy-based learning, 3) policy-based learning and supervised learning and 4) reinforcement learning and imitation learning.
****
Shifted and Squeezed 8-bit Floating Point format for Low-Precision Training of Deep Neural Networks
Leopold Cambier,Anahita Bhiwandiwalla,Ting Gong,Oguz H. Elibol,Mehran Nekuii,Hanlin Tang
Training with larger number of parameters while keeping fast iterations is an increasingly
adopted strategy and trend for developing better performing Deep Neural
Network (DNN) models. This necessitates increased memory footprint and
computational requirements for training. Here we introduce a novel methodology
for training deep neural networks using 8-bit floating point (FP8) numbers.
Reduced bit precision allows for a larger effective memory and increased computational
speed. We name this method Shifted and Squeezed FP8 (S2FP8). We
show that, unlike previous 8-bit precision training methods, the proposed method

works out of the box for representative models: ResNet50, Transformer and NCF. The method can maintain model accuracy without requiring fine-tuning loss scaling
parameters or keeping certain layers in single precision. We introduce two learnable statistics of the DNN tensors - shifted and squeezed factors that are used
to optimally adjust the range of the tensors in 8-bits, thus minimizing the loss in
information due to quantization.
****
Intriguing Properties of Adversarial Training at Scale
Cihang Xie,Alan Yuille
Adversarial training is one of the main defenses against adversarial attacks. In this paper, we provide the first rigorous study on diagnosing elements of large-scale adversarial training on ImageNet, which reveals two intriguing properties.

First, we study the role of normalization. Batch normalization (BN) is a crucial element for achieving state-of-the-art performance on many vision tasks, but we show it may prevent networks from obtaining strong robustness in adversarial training. One unexpected observation is that, for models trained with BN, simply removing clean images from training data largely boosts adversarial robustness, i.e., 18.3%. We relate this phenomenon to the hypothesis that clean images and adversarial images are drawn from two different domains. This two-domain hypothesis may explain the issue of BN when training with a mixture of clean and adversarial images, as estimating normalization statistics of this mixture distribution is challenging. Guided by this two-domain hypothesis, we show disentangling the mixture distribution for normalization, i.e., applying separate BNs to clean and adversarial images for statistics estimation, achieves much stronger robustness. Additionally, we find that enforcing BNs to behave consistently at training and testing can further enhance robustness.

Second, we study the role of network capacity. We find our so-called "deep" networks are still shallow for the task of adversarial learning. Unlike traditional classification tasks where accuracy is only marginally improved by adding more layers to "deep" networks (e.g., ResNet-152), adversarial training exhibits a much stronger demand on deeper networks to achieve higher adversarial robustness. This robustness improvement can be observed substantially and consistently even by pushing the network capacity to an unprecedented scale, i.e., ResNet-638.
****
Point Process Flows
Nazanin Mehrasa,Ruizhi Deng,Mohamed Osama Ahmed,Bo Chang,Jiawei He,Thibaut Durand,Marcus Brubaker,Greg Mori
Event sequences can be modeled by temporal point processes (TPPs) to capture their asynchronous and probabilistic nature. We propose an intensity-free framework that directly models the point process as a non-parametric distribution by utilizing normalizing flows. This approach is capable of capturing highly complex temporal distributions and does not rely on restrictive parametric forms. Comparisons with state-of-the-art baseline models on both synthetic and challenging real-life datasets show that the proposed framework is effective at modeling the stochasticity of discrete event sequences.
****
Cover Filtration and Stable Paths in the Mapper
Dustin L. Arendt,Matthew Broussard,Bala Krishnamoorthy,Nathaniel Saul
The contributions of this paper are two-fold. We define a new filtration called the cover filtration built from a single cover based on a generalized Steinhaus distance, which is a generalization of Jaccard distance. We then develop a language and theory for stable paths within this filtration, inspired by ideas of persistent homology. This framework can be used to develop several new learning representations in applications where an obvious metric may not be defined but a co

ver is readily available. We demonstrate the utility of our framework as applied to recommendation systems and explainable machine learning.

We demonstrate a new perspective for modeling recommendation system data sets that does not require manufacturing a bespoke metric. As a direct application, we find that the stable paths identified by our framework in a movies data set represent a sequence of movies constituting a gentle transition and ordering from one genre to another.

For explainable machine learning, we apply the Mapper for model induction, providing explanations in the form of paths between subpopulations. Our framework provides an alternative way of building a filtration from a single mapper that is then used to explore stable paths. As a direct illustration, we build a mapper from a supervised machine learning model trained on the FashionMNIST data set. We show that the stable paths in the cover filtration provide improved explanations of relationships between subpopulations of images.

****

Fully Polynomial-Time Randomized Approximation Schemes for Global Optimization of High-Dimensional Folded Concave Penalized Generalized Linear Models
Charles Hernandez,Hungyi Lee,Hongchen Liu
Global solutions to high-dimensional sparse estimation problems with a folded concave penalty (FCP) have been shown to be statistically desirable but are strongly NP-hard to compute, which implies the non-existence of a pseudo-polynomial time global optimization schemes in the worst case. This paper shows that, with high probability, a global solution to the formulation for a FCP-based high-dimensional generalized linear model coincides with a stationary point characterized by the significant subspace second order necessary conditions (S^{3}ONC). Since the desired S^{3}ONC solution admits a fully polynomial-time approximation schemes (FPTAS), we thus have shown the existence of fully polynomial-time randomized approximation scheme (FPRAS) for a strongly NP-hard problem. We further demonstrate two versions of the FPRAS for generating the desired S^{3}ONC solutions. One follows the paradigm of an interior point trust region algorithm and the other is the well-studied local linear approximation (LLA). Our analysis thus provides new techniques for global optimization of certain NP-Hard problems and new insights on the effectiveness of LLA.
****

Learning Neural Surrogate Model for Warm-Starting Bayesian Optimization
Haotian Zhang,Jian Sun,Zongben Xu
Bayesian optimization is an effective tool to optimize black-box functions and popular for hyper-parameter tuning in machine learning. Traditional Bayesian optimization methods are based on Gaussian process (GP), relying on a GP-based surrogate model for sampling points of the function of interest. In this work, we consider transferring knowledge from related problems to target problem by learning an initial surrogate model for warm-starting Bayesian optimization. We propose a neural network-based surrogate model to estimate the function mean value in GP. Then we design a novel weighted Reptile algorithm with sampling strategy to learn an initial surrogate model from meta train set. The initial surrogate model is learned to be able to well adapt to new tasks. Extensive experiments show that this warm-starting technique enables us to find better minimizer or hyper-parameters than traditional GP and previous warm-starting methods.
****

Scalable Differentially Private Data Generation via Private Aggregation of Teacher Ensembles
Yunhui Long,Suxin Lin,Zhuolin Yang,Carl A. Gunter,Han Liu,Bo Li
We present a novel approach named G-PATE for training differentially private data generator. The generator can be used to produce synthetic datasets with strong privacy guarantee while preserving high data utility. Our approach leverages generative adversarial nets to generate data and exploits the PATE (Private Aggregation of Teacher Ensembles) framework to protect data privacy. Compared to exis

ting methods, our approach significantly improves the use of privacy budget. Th is is possible since we only need to ensure differential privacy for the genera tor, which is the part of the model that actually needs to be published for priv ate data generation. To achieve this, we connect a student generator with an en semble of teacher discriminators and propose a private gradient aggregation mech anism to ensure differential privacy on all the information that flows from the teacher discriminators to the student generator. Theoretically, we prove that ou r algorithm ensures differential privacy for the generator. Empirically, we pr ovide thorough experiments to demonstrate the superiority of our method over pri or work on both image and non-image datasets.
****

Knowledge Graph Embedding: A Probabilistic Perspective and Generalization Bounds
Ondrej Kuzelka,Yuyi Wang
We study theoretical properties of embedding methods for knowledge graph complet ion under the missing completely at random assumption. We prove generalization e rror bounds for this setting. Even though the missing completely at random setti ng may seem naive, it is actually how knowledge graph embedding methods are typi cally benchmarked in the literature. Our results provide, to certain extent, an explanation for why knowledge graph embedding methods work (as much as classical learning theory results provide explanations for classical learning from i.i.d. data).
****

Adversarial Paritial Multi-label Learning
Yan Yan,Yuhong Guo
Partial multi-label learning (PML), which tackles the problem of learning multi-label prediction models from instances with overcomplete noisy annotations, has recently started gaining attention from the research community. In this paper, w e propose a novel adversarial learning model, PML-GAN, under a generalized encod er-decoder framework for partial multi-label learning. The PML-GAN model uses a disambiguation network to identify noisy labels and uses a multi-label predictio n network to map the training instances to the disambiguated label vectors, whil e deploying a generative adversarial network as an inverse mapping from label ve ctors to data samples in the input feature space. The learning of the overall mo del corresponds to a minimax adversarial game, which enhances the correspondence of input features with the output labels. Extensive experiments are conducted o n multiple datasets, while the proposed model demonstrates the state-of-the-art performance for partial multi-label learning.
****

Adversarial Interpolation Training: A Simple Approach for Improving Model Robust ness
Haichao Zhang,Wei Xu
We propose a simple approach for adversarial training. The proposed approach uti lizes an adversarial interpolation scheme for generating adversarial images and accompanying adversarial labels, which are then used in place of the original da ta for model training. The proposed approach is intuitive to understand, simple to implement and achieves state-of-the-art performance. We evaluate the proposed approach on a number of datasets including CIFAR10, CIFAR100 and SVHN. Extensiv e empirical results compared with several state-of-the-art methods against diffe rent attacks verify the effectiveness of the proposed approach.
****

Agent as Scientist: Learning to Verify Hypotheses
Kenneth Marino,Rob Fergus,Arthur Szlam,Abhinav Gupta
In this paper, we formulate hypothesis verification as a reinforcement learning problem. Specifically, we aim to build an agent that, given a hypothesis about t he dynamics of the world can take actions to generate observations which can hel p predict whether the hypothesis is true or false. Our first observation is that agents trained end-to-end with the reward fail to learn to solve this problem. In order to train the agents, we exploit the underlying structure in the majorit y of hypotheses -- they can be formulated as triplets (pre-condition, action seq uence, post-condition). Once the agents have been pretrained to verify hypothes

es with this structure, they can be fine-tuned to verify more general hypotheses . Our work takes a step towards a ``scientist agent'' that develops an understanding of the world by generating and testing hypotheses about its environment.
****

CRNet: Image Super-Resolution Using A Convolutional Sparse Coding Inspired Network

Menglei Zhang,Zhou Liu,Jingwei He,Lei Yu

Convolutional Sparse Coding (CSC) has been attracting more and more attention in recent years, for making full use of image global correlation to improve performance on various computer vision applications. However, very few studies focus on solving CSC based image Super-Resolution (SR) problem. As a consequence, there is no significant progress in this area over a period of time. In this paper, we exploit the natural connection between CSC and Convolutional Neural Networks (CNN) to address CSC based image SR. Specifically, Convolutional Iterative Soft Thresholding Algorithm (CISTA) is introduced to solve CSC problem and it can be implemented using CNN architectures. Then we develop a novel CSC based SR framework analogy to the traditional SC based SR methods. Two models inspired by this framework are proposed for pre-/post-upsampling SR, respectively. Compared with recent state-of-the-art SR methods, both of our proposed models show superior performance in terms of both quantitative and qualitative measurements.
****

Deep Double Descent: Where Bigger Models and More Data Hurt

Preetum Nakkiran,Gal Kaplun,Yamini Bansal,Tristan Yang,Boaz Barak,Ilya Sutskever

We show that a variety of modern deep learning tasks exhibit a "double-descent" phenomenon where, as we increase model size, performance first gets worse and then gets better. Moreover, we show that double descent occurs not just as a function of model size, but also as a function of the number of training epochs. We unify the above phenomena by defining a new complexity measure we call the effective model complexity, and conjecture a generalized double descent with respect to this measure. Furthermore, our notion of model complexity allows us to identify certain regimes where increasing (even quadrupling) the number of train samples actually hurts test performance.
****

Multigrid Neural Memory

Tri Huynh,Michael Maire,Matthew R. Walter

We introduce a novel architecture that integrates a large addressable memory space into the core functionality of a deep neural network. Our design distributes both memory addressing operations and storage capacity over many network layers. Distinct from strategies that connect neural networks to external memory banks, our approach co-locates memory with computation throughout the network structure. Mirroring recent architectural innovations in convolutional networks, we organize memory into a multiresolution hierarchy, whose internal connectivity enables learning of dynamic information routing strategies and data-dependent read/write operations. This multigrid spatial layout permits parameter-efficient scaling of memory size, allowing us to experiment with memories substantially larger than those in prior work. We demonstrate this capability on synthetic exploration and mapping tasks, where the network is able to self-organize and retain long-term memory for trajectories of thousands of time steps. On tasks decoupled from any notion of spatial geometry, such as sorting or associative recall, our design functions as a truly generic memory and yields results competitive with those of the recently proposed Differentiable Neural Computer.
****

ASGen: Answer-containing Sentence Generation to Pre-Train Question Generator for Scale-up Data in Question Answering

Akhil Kedia,Sai Chetan Chinthakindi,Seohyun Back,Haejun Lee,Jaegul Choo

Numerous machine reading comprehension (MRC) datasets often involve manual annotation, requiring enormous human effort, and hence the size of the dataset remains significantly smaller than the size of the data available for unsupervised learning. Recently, researchers proposed a model for generating synthetic question-and-answer data from large corpora such as Wikipedia. This model is utilized to

generate synthetic data for training an MRC model before fine-tuning it using th
e original MRC dataset. This technique shows better performance than other gener
al pre-training techniques such as language modeling, because the characteristic
s of the generated data are similar to those of the downstream MRC data. However
, it is difficult to have high-quality synthetic data comparable to human-annota
ted MRC datasets. To address this issue, we propose Answer-containing Sentence G
eneration (ASGen), a novel pre-training method for generating synthetic data inv
olving two advanced techniques, (1) dynamically determining K answers and (2) pr
e-training the question generator on the answer-containing sentence generation t
ask. We evaluate the question generation capability of our method by comparing t
he BLEU score with existing methods and test our method by fine-tuning the MRC m
odel on the downstream MRC data after training on synthetic data. Experimental r
esults show that our approach outperforms existing generation methods and increa
ses the performance of the state-of-the-art MRC models across a range of MRC dat
asets such as SQuAD-v1.1, SQuAD-v2.0, KorQuAD and QUASAR-T without any architect
ural modifications to the original MRC model.
****

Distribution-Guided Local Explanation for Black-Box Classifiers
Weijie Fu,Meng Wang,Mengnan Du,Ninghao Liu,Shijie Hao,Xia Hu
Existing local explanation methods provide an explanation for each decision of b
lack-box classifiers, in the form of relevance scores of features according to t
heir contributions. To obtain satisfying explainability, many methods introduce
ad hoc constraints into the classification loss to regularize these relevance sc
ores. However, the large information gap between the classification loss and the
se constraints increases the difficulty of tuning hyper-parameters. To bridge th
is gap, in this paper we present a simple but effective mask predictor. Specific
ally, we model the above constraints with a distribution controller, and integra
te it with a neural network to directly guide the distribution of relevance scor
es. The benefit of this strategy is to facilitate the setting of involved hyper-
parameters, and enable discriminative scores over supporting features. The exper
imental results demonstrate that our method outperforms others in terms of faith
fulness and explainability. Meanwhile, it also provides effective saliency maps
for explaining each decision.
****

Decoding As Dynamic Programming For Recurrent Autoregressive Models
Najam Zaidi,Trevor Cohn,Gholamreza Haffari
Decoding in autoregressive models (ARMs) consists of searching for a high scorin
g output sequence under the trained model. Standard decoding methods, based on
unidirectional greedy algorithm or beam search, are suboptimal due to error prop
agation and myopic decisions which do not account for future steps in the genera
tion process. In this paper we present a novel decoding approach based on the me
thod of auxiliary coordinates (Carreira-Perpinan & Wang, 2014) to address the af
orementioned shortcomings. Our method introduces discrete variables for output
tokens, and auxiliary continuous variables representing the states of the under
lying ARM. The auxiliary variables lead to a factor graph approximation of the A
RM, whose maximum a posteriori (MAP) inference is found exactly using dynamic pr
ogramming. The MAP inference is then used to recreate an improved factor graph a
pproximation of the ARM via updated auxiliary variables. We then extend our appr
oach to decode in an ensemble of ARMs, possibly with different generation orders
, which is out of reach for the standard unidirectional decoding algorithms. Ex
periments on the text infilling task over SWAG and Daily Dialogue datasets show
that our decoding method is superior to strong unidirectional decoding baselines
.
****

Compressed Sensing with Deep Image Prior and Learned Regularization
Dave Van Veen,Ajil Jalal,Mahdi Soltanolkotabi,Eric Price,Sriram Vishwanath,Alexa
ndros G. Dimakis
We propose a novel method for compressed sensing recovery using
untrained deep generative models. Our method is based on the recently
proposed Deep Image Prior (DIP), wherein the convolutional weights of

the network are optimized to match the observed measurements. We show
that this approach can be applied to solve any differentiable linear inverse
problem, outperforming previous unlearned methods. Unlike various learned approa
ches based on generative models, our method does not require pre-training over l
arge datasets. We further introduce a novel learned regularization technique, wh
ich incorporates prior information on the network weights. This reduces reconstr
uction error, especially for noisy measurements. Finally we prove that, using th
e DIP optimization approach, moderately overparameterized single-layer networks
trained can perfectly fit any signal despite the nonconvex nature of the fitting
 problem. This theoretical result provides justification for early stopping.
****

Gradient Surgery for Multi-Task Learning
Tianhe Yu,Saurabh Kumar,Abhishek Gupta,Karol Hausman,Sergey Levine,Chelsea Finn
While deep learning and deep reinforcement learning systems have demonstrated im
pressive results in domains such as image classification, game playing, and robo
tic control, data efficiency remains a major challenge, particularly as these al
gorithms learn individual tasks from scratch. Multi-task learning has emerged as
 a promising approach for sharing structure across multiple tasks to enable more
 efficient learning. However, the multi-task setting presents a number of optimi
zation challenges, making it difficult to realize large efficiency gains compare
d to learning tasks independently. The reasons why multi-task learning is so cha
llenging compared to single task learning are not fully understood. Motivated by
 the insight that gradient interference causes optimization challenges, we devel
op a simple and general approach for avoiding interference between gradients fro
m different tasks, by altering the gradients through a technique we refer to as
“gradient surgery”. We propose a form of gradient surgery that projects the grad
ient of a task onto the normal plane of the gradient of any other task that has
a conflicting gradient. On a series of challenging multi-task supervised and mul
ti-task reinforcement learning problems, we find that this approach leads to sub
stantial gains in efficiency and performance. Further, it can be effectively co
mbined with previously-proposed multi-task architectures for enhanced performanc
e in a model-agnostic way.
****

SINGLE PATH ONE-SHOT NEURAL ARCHITECTURE SEARCH WITH UNIFORM SAMPLING
Zichao Guo,Xiangyu Zhang,Haoyuan Mu,Wen Heng,Zechun Liu,Yichen Wei,Jian Sun
We revisit the one-shot Neural Architecture Search (NAS) paradigm and analyze it
s advantages over existing NAS approaches. Existing one-shot method (Benderet al
., 2018), however, is hard to train and not yet effective on large scale dataset
s like ImageNet. This work propose a Single Path One-Shot model to address the
challenge in the training. Our central idea is to construct a simplified supern
et, where all architectures are single paths so that weight co-adaption problem
is alleviated. Training is performed by uniform path sampling. All architectures
 (and their weights) are trained fully and equally.
Comprehensive experiments verify that our approach is flexible and effective. I
t is easy to train and fast to search. It effortlessly supports complex search
spaces(e.g., building blocks, channel, mixed-precision quantization) and differe
nt search constraints (e.g., FLOPs, latency). It is thus convenient to use for
various needs. It achieves start-of-the-art performance on the large dataset Ima
geNet.
****

Synthesizing Programmatic Policies that Inductively Generalize
Jeevana Priya Inala,Osbert Bastani,Zenna Tavares,Armando Solar-Lezama
Deep reinforcement learning has successfully solved a number of challenging cont
rol tasks. However, learned policies typically have difficulty generalizing to n
ovel environments. We propose an algorithm for learning programmatic state machi
ne policies that can capture repeating behaviors. By doing so, they have the abi
lity to generalize to instances requiring an arbitrary number of repetitions, a
property we call inductive generalization. However, state machine policies are h
ard to learn since they consist of a combination of continuous and discrete stru
ctures. We propose a learning framework called adaptive teaching, which learns a

state machine policy by imitating a teacher; in contrast to traditional imitation learning, our teacher adaptively updates itself based on the structure of the student. We show that our algorithm can be used to learn policies that inductively generalize to novel environments, whereas traditional neural network policies fail to do so.
****
Transformer-XH: Multi-Evidence Reasoning with eXtra Hop Attention
Chen Zhao,Chenyan Xiong,Corby Rosset,Xia Song,Paul Bennett,Saurabh Tiwary
Transformers have achieved new heights modeling natural language as a sequence of text tokens. However, in many real world scenarios, textual data inherently exhibits structures beyond a linear sequence such as trees and graphs; many tasks require reasoning with evidence scattered across multiple pieces of texts. This paper presents Transformer-XH, which uses eXtra Hop attention to enable intrinsic modeling of structured texts in a fully data-driven way. Its new attention mechanism naturally “hops” across the connected text sequences in addition to attending over tokens within each sequence. Thus, Transformer-XH better conducts joint multi-evidence reasoning by propagating information between documents and constructing global contextualized representations. On multi-hop question answering, Transformer-XH leads to a simpler multi-hop QA system which outperforms previous state-of-the-art on the HotpotQA FullWiki setting. On FEVER fact verification, applying Transformer-XH provides state-of-the-art accuracy and excels on claims whose verification requires multiple evidence.
****
Variational Hyper RNN for Sequence Modeling
Ruizhi Deng,Yanshuai Cao,Bo Chang,Leonid Sigal,Greg Mori,Marcus Brubaker
In this work, we propose a novel probabilistic sequence model that excels at capturing high variability in time series data, both across sequences and within an individual sequence. Our method uses temporal latent variables to capture information about the underlying data pattern and dynamically decodes the latent information into modifications of weights of the base decoder and recurrent model. The efficacy of the proposed method is demonstrated on a range of synthetic and real-world sequential data that exhibit large scale variations, regime shifts, and complex dynamics.
****
Generalization through Memorization: Nearest Neighbor Language Models
Urvashi Khandelwal,Omer Levy,Dan Jurafsky,Luke Zettlemoyer,Mike Lewis
We introduce kNN-LMs, which extend a pre-trained neural language model (LM) by linearly interpolating it with a k-nearest neighbors (kNN) model. The nearest neighbors are computed according to distance in the pre-trained LM embedding space, and can be drawn from any text collection, including the original LM training data. Applying this transformation to a strong Wikitext-103 LM, with neighbors drawn from the original training set, our kNN-LM achieves a new state-of-the-art perplexity of 15.79 -- a 2.9 point improvement with no additional training. We also show that this approach has implications for efficiently scaling up to larger training sets and allows for effective domain adaptation, by simply varying the nearest neighbor datastore, again without further training. Qualitatively, the model is particularly helpful in predicting rare patterns, such as factual knowledge. Together, these results strongly suggest that learning similarity between sequences of text is easier than predicting the next word, and that nearest neighbor search is an effective approach for language modeling in the long tail.
****
Comparing Rewinding and Fine-tuning in Neural Network Pruning
Alex Renda,Jonathan Frankle,Michael Carbin
Many neural network pruning algorithms proceed in three steps: train the network to completion, remove unwanted structure to compress the network, and retrain the remaining structure to recover lost accuracy. The standard retraining technique, fine-tuning, trains the unpruned weights from their final trained values using a small fixed learning rate. In this paper, we compare fine-tuning to alternative retraining techniques. Weight rewinding (as proposed by Frankle et al., (20

19)), rewinds unpruned weights to their values from earlier in training and retr
ains them from there using the original training schedule. Learning rate rewindi
ng (which we propose) trains the unpruned weights from their final values using
the same learning rate schedule as weight rewinding. Both rewinding techniques o
utperform fine-tuning, forming the basis of a network-agnostic pruning algorithm
that matches the accuracy and compression ratios of several more network-specif
ic state-of-the-art techniques.

****
Simple is Better: Training an End-to-end Contract Bridge Bidding Agent without H
uman Knowledge
Qucheng Gong,Yu Jiang,Yuandong Tian
Contract bridge is a multi-player imperfect-information game where one partnersh
ip collaborate with each other to compete against the other partnership. The gam
e consists of two phases: bidding and playing. While playing is relatively easy
for modern software, bidding is challenging and requires agents to learn a commu
nication protocol to reach the optimal contract jointly, with their own private
information. The agents need to exchange information to their partners, and inte
rfere opponents, through a sequence of actions. In this work, we train a strong
agent to bid competitive bridge purely through selfplay, outperforming WBridge5,
a championship-winning software. Furthermore, we show that explicitly modeling
belief is not necessary in boosting the performance. To our knowledge, this is t
he first competitive bridge agent that is trained with no domain knowledge. It o
utperforms previous state-of-the-art that use human replays with 70x fewer numbe
r of parameters.
****
The Sooner The Better: Investigating Structure of Early Winning Lottery Tickets
Shihui Yin,Kyu-Hyoun Kim,Jinwook Oh,Naigang Wang,Mauricio Serrano,Jae-Sun Seo,Ju
ngwook Choi
The recent success of the lottery ticket hypothesis by Frankle & Carbin (2018) s
uggests that small, sparsified neural networks can be trained as long as the net
work is initialized properly. Several follow-up discussions on the initializatio
n of the sparsified model have discovered interesting characteristics such as th
e necessity of rewinding (Frankle et al. (2019)), importance of sign of the init
ial weights (Zhou et al. (2019)), and the transferability of the winning lottery
tickets (S. Morcos et al. (2019)). In contrast, another essential aspect of the
winning ticket, the structure of the sparsified model, has been little discusse
d. To find the lottery ticket, unfortunately, all the prior work still relies on
computationally expensive iterative pruning.

In this work, we conduct an in-depth investigation of the structure of winning l
ottery tickets. Interestingly, we discover that there exist many lottery tickets
that can achieve equally good accuracy much before the regular training schedul
e even finishes. We provide insights into the structure of these early winning t
ickets with supporting evidence. 1) Under stochastic gradient descent optimizati
on, lottery ticket emerges when weight magnitude of a model saturates; 2) Prunin
g before the saturation of a model causes the loss of capability in learning com
plex patterns, resulting in the accuracy degradation. We employ the memorization
capacity analysis to quantitatively confirm it, and further explain why gradual
pruning can achieve better accuracy over the one-shot pruning. Based on these i
nsights, we discover the early winning tickets for various ResNet architectures
on both CIFAR10 and ImageNet, achieving state-of-the-art accuracy at a high prun
ing rate without expensive iterative pruning. In the case of ResNet50 on ImageNe
t, this comes to the winning ticket of 75:02% Top-1 accuracy at 80% pruning rate
in only 22% of the total epochs for iterative pruning.
****
Long History Short-Term Memory for Long-Term Video Prediction
Wonmin Byeon,Jan Kautz
While video prediction approaches have advanced considerably in recent years, le
arning to predict long-term future is challenging — ambiguous future or error pr

opagation over time yield blurry predictions. To address this challenge, existin g algorithms rely on extra supervision (e.g., action or object pose), motion flo w learning, or adversarial training. In this paper, we propose a new recurrent u nit, Long History Short-Term Memory (LH-STM). LH-STM incorporates long history s tates into a recurrent unit to learn longer range dependencies. To capture spati o-temporal dynamics in videos, we combined LH-STM with the Context-aware Video P rediction model (ContextVP). Our experiments on the KTH human actions and BAIR r obot pushing datasets demonstrate that our approach produces not only sharper ne ar-future predictions, but also farther into the future compared to the state-of -the-art methods.

****

Adversarial training with perturbation generator networks

Hyeungill Lee,Sungyeob Han,Jungwoo Lee

Despite the remarkable development of recent deep learning techniques, neural ne tworks are still vulnerable to adversarial attacks, i.e., methods that fool the neural networks with perturbations that are too small for human eyes to perceive . Many adversarial training methods were introduced as to solve this problem, us ing adversarial examples as a training data. However, these adversarial attack m ethods used in these techniques are fixed, making the model stronger only to att acks used in training, which is widely known as an overfitting problem. In this paper, we suggest a novel adversarial training approach. In addition to the clas sifier, our method adds another neural network that generates the most effective adversarial perturbation by finding the weakness of the classifier. This pertur bation generator network is trained to produce perturbations that maximize the l oss function of the classifier, and these adversarial examples train the classif ier with a true label. In short, the two networks compete with each other, perfo rming a minimax game. In this scenario, attack patterns created by the generator network are adaptively altered to the classifier, mitigating the overfitting pr oblem mentioned above. We theoretically proved that our minimax optimization pro blem is equivalent to minimizing the adversarial loss after all. Beyond this, we proposed an evaluation method that could accurately compare a wide-range of adv ersarial algorithms. Experiments with various datasets show that our method outp erforms conventional adversarial algorithms.

****

Single Episode Policy Transfer in Reinforcement Learning

Jiachen Yang,Brenden Petersen,Hongyuan Zha,Daniel Faissol

Transfer and adaptation to new unknown environmental dynamics is a key challenge for reinforcement learning (RL). An even greater challenge is performing near-o ptimally in a single attempt at test time, possibly without access to dense rewa rds, which is not addressed by current methods that require multiple experience rollouts for adaptation. To achieve single episode transfer in a family of envir onments with related dynamics, we propose a general algorithm that optimizes a p robe and an inference model to rapidly estimate underlying latent variables of t est dynamics, which are then immediately used as input to a universal control po licy. This modular approach enables integration of state-of-the-art algorithms f or variational inference or RL. Moreover, our approach does not require access t o rewards at test time, allowing it to perform in settings where existing adapti ve approaches cannot. In diverse experimental domains with a single episode test constraint, our method significantly outperforms existing adaptive approaches a nd shows favorable performance against baselines for robust transfer.

****

Inducing Stronger Object Representations in Deep Visual Trackers

Ross Goroshin,Jonathan Tompson,Debidatta Dwibedi

Fully convolutional deep correlation networks are integral components of state-o f-
the-art approaches to single object visual tracking. It is commonly assumed that these networks perform tracking by detection by matching features of the object instance with features of the entire frame. Strong architectural priors and cond itioning
on the object representation is thought to encourage this tracking strategy.

Despite these strong priors, we show that deep trackers often default to “tracking-
by-saliency” detection – without relying on the object instance representation. Our
analysis shows that despite being a useful prior, salience detection can prevent the
emergence of more robust tracking strategies in deep networks. This leads us to
introduce an auxiliary detection task that encourages more discriminative object
representations that improve tracking performance.
****

Towards Stabilizing Batch Statistics in Backward Propagation of Batch Normalization
Junjie Yan,Ruosi Wan,Xiangyu Zhang,Wei Zhang,Yichen Wei,Jian Sun
Batch Normalization (BN) is one of the most widely used techniques in Deep Learning field. But its performance can awfully degrade with insufficient batch size. This weakness limits the usage of BN on many computer vision tasks like detection or segmentation, where batch size is usually small due to the constraint of memory consumption. Therefore many modified normalization techniques have been proposed, which either fail to restore the performance of BN completely, or have to introduce additional nonlinear operations in inference procedure and increase huge consumption. In this paper, we reveal that there are two extra batch statistics involved in backward propagation of BN, on which has never been well discussed before. The extra batch statistics associated with gradients also can severely affect the training of deep neural network. Based on our analysis, we propose a novel normalization method, named Moving Average Batch Normalization (MABN). MABN can completely restore the performance of vanilla BN in small batch cases, without introducing any additional nonlinear operations in inference procedure. We prove the benefits of MABN by both theoretical analysis and experiments. Our experiments demonstrate the effectiveness of MABN in multiple computer vision tasks including ImageNet and COCO. The code has been released in https://github.com/megvii-model/MABN.
****

STABILITY AND CONVERGENCE THEORY FOR LEARNING RESNET: A FULL CHARACTERIZATION
Huishuai Zhang,Da Yu,Mingyang Yi,Wei Chen,Tie-yan Liu
ResNet structure has achieved great success since its debut. In this paper, we study the stability of learning ResNet. Specifically, we consider the ResNet block $h_l = \phi(h_{l-1}+\tau\cdot g(h_{l-1}))$ where $\phi(\cdot)$ is ReLU activation and τ is a scalar. We show that for standard initialization used in practice, $\tau =1/\Omega(\sqrt{L})$ is a sharp value in characterizing the stability of forward/backward process of ResNet, where L is the number of residual blocks. Specifically, stability is guaranteed for $\tau\le 1/\Omega(\sqrt{L})$ while conversely forward process explodes when $\tau>L^{-\frac{1}{2}+c}$ for a positive constant c. Moreover, if ResNet is properly over-parameterized, we show for $\tau \le 1/\tilde{\Omega}(\sqrt{L})$ gradient descent is guaranteed to find the global minima \footnote{We use $\tilde{\Omega}(\cdot)$ to hide logarithmic factor.}, which significantly enlarges the range of $\tau\le 1/\tilde{\Omega}(L)$ that admits global convergence in previous work. We also demonstrate that the over-parameterization requirement of ResNet only weakly depends on the depth, which corroborates the advantage of ResNet over vanilla feedforward network. Empirically, with $\tau\le1/\sqrt{L}$, deep ResNet can be easily trained even without normalization layer. Moreover, adding $\tau=1/\sqrt{L}$ can also improve the performance of ResNet with normalization layer.
****

Training Deep Neural Networks with Partially Adaptive Momentum
Jinghui Chen,Dongruo Zhou,Yiqi Tang,Ziyan Yang,Yuan Cao,Quanquan Gu
Adaptive gradient methods, which adopt historical gradient information to automatically adjust the learning rate, despite the nice property of fast convergence, have been observed to generalize worse than stochastic gradient descent (SGD) with momentum in training deep neural networks. This leaves how to close the generalization gap of adaptive gradient methods an open problem. In this work, we sh

ow that adaptive gradient methods such as Adam, Amsgrad, are sometimes ``over ad
apted''. We design a new algorithm, called Partially adaptive momentum estimatio
n method, which unifies the Adam/Amsgrad with SGD by introducing a partial adapt
ive parameter p, to achieve the best from both worlds. We also prove the conve
rgence rate of our proposed algorithm to a stationary point in the stochastic no
nconvex optimization setting. Experiments on standard benchmarks show that our p
roposed algorithm can maintain fast convergence rate as Adam/Amsgrad while gener
alizing as well as SGD in training deep neural networks. These results would sug
gest practitioners pick up adaptive gradient methods once again for faster train
ing of deep neural networks.
****

NeurQuRI: Neural Question Requirement Inspector for Answerability Prediction in
Machine Reading Comprehension
Seohyun Back,Sai Chetan Chinthakindi,Akhil Kedia,Haejun Lee,Jaegul Choo
Real-world question answering systems often retrieve potentially relevant docume
nts to a given question through a keyword search, followed by a machine reading
comprehension (MRC) step to find the exact answer from them. In this process, it
 is essential to properly determine whether an answer to the question exists in
a given document. This task often becomes complicated when the question involves
 multiple different conditions or requirements which are to be met in the answer
. For example, in a question "What was the projection of sea level increases in
the fourth assessment report?", the answer should properly satisfy several condi
tions, such as "increases" (but not decreases) and "fourth" (but not third). To
address this, we propose a neural question requirement inspection model called N
eurQuRI that extracts a list of conditions from the question, each of which shou
ld be satisfied by the candidate answer generated by an MRC model. To check whet
her each condition is met, we propose a novel, attention-based loss function. We
 evaluate our approach on SQuAD 2.0 dataset by integrating the proposed module w
ith various MRC models, demonstrating the consistent performance improvements ac
ross a wide range of state-of-the-art methods.
****

Learning Latent Representations for Inverse Dynamics using Generalized Experienc
es
Aditi Mavalankar,Sicun Gao
Many practical robot locomotion tasks require agents to use control policies tha
t can be parameterized by goals. Popular deep reinforcement learning approaches
in this direction involve learning goal-conditioned policies or value functions,
 or Inverse Dynamics Models (IDMs). IDMs map an agent’s current state and desire
d goal to the required actions. We show that the key to achieving good performan
ce with IDMs lies in learning the information shared between equivalent experien
ces, so that they can be generalized to unseen scenarios. We design a training p
rocess that guides the learning of latent representations to encode this shared
information. Using a limited number of environment interactions, our agent is ab
le to efficiently navigate to arbitrary points in the goal space. We demonstrate
 the effectiveness of our approach in high-dimensional locomotion environments s
uch as the Mujoco Ant, PyBullet Humanoid, and PyBullet Minitaur. We provide quan
titative and qualitative results to show that our method clearly outperforms com
peting baseline approaches.
****

Learning The Difference That Makes A Difference With Counterfactually-Augmented
Data
Divyansh Kaushik,Eduard Hovy,Zachary Lipton
Despite alarm over the reliance of machine learning systems on so-called spuriou
s patterns, the term lacks coherent meaning in standard statistical frameworks.
However, the language of causality offers clarity: spurious associations are due
 to confounding (e.g., a common cause), but not direct or indirect causal effect
s. In this paper, we focus on natural language processing, introducing methods a
nd resources for training models less sensitive to spurious patterns. Given doc
uments and their initial labels, we task humans with revising each document so t
hat it (i) accords with a counterfactual target label; (ii) retains internal coh

erence; and (iii) avoids unnecessary changes. Interestingly, on sentiment analysis and natural language inference tasks, classifiers trained on original data fail on their counterfactually-revised counterparts and vice versa. Classifiers trained on combined datasets perform remarkably well, just shy of those specialized to either domain. While classifiers trained on either original or manipulated data alone are sensitive to spurious features (e.g., mentions of genre), models trained on the combined data are less sensitive to this signal. Both datasets are publicly available.
****

Differentiable Architecture Compression
Shashank Singh,Ashish Khetan,Zohar Karnin
In many learning situations, resources at inference time are significantly more constrained than resources at training time. This paper studies a general paradigm, called Differentiable ARchitecture Compression (DARC), that combines model compression and architecture search to learn models that are resource-efficient at inference time. Given a resource-intensive base architecture, DARC utilizes the training data to learn which sub-components can be replaced by cheaper alternatives. The high-level technique can be applied to any neural architecture, and we report experiments on state-of-the-art convolutional neural networks for image classification. For a WideResNet with 97.2% accuracy on CIFAR-10, we improve single-sample inference speed by 2.28X and memory footprint by 5.64X, with no accuracy loss. For a ResNet with 79.15% Top-1 accuracy on ImageNet, we improve batch inference speed by 1.29X and memory footprint by 3.57X with 1% accuracy loss. We also give theoretical Rademacher complexity bounds in simplified cases, showing how DARC avoids over-fitting despite over-parameterization.
****

The Early Phase of Neural Network Training
Jonathan Frankle,David J. Schwab,Ari S. Morcos
Recent studies have shown that many important aspects of neural network learning take place within the very earliest iterations or epochs of training. For example, sparse, trainable sub-networks emerge (Frankle et al., 2019), gradient descent moves into a small subspace (Gur-Ari et al., 2018), and the network undergoes a critical period (Achille et al., 2019). Here we examine the changes that deep neural networks undergo during this early phase of training. We perform extensive measurements of the network state and its updates during these early iterations of training, and leverage the framework of Frankle et al. (2019) to quantitatively probe the weight distribution and its reliance on various aspects of the dataset. We find that, within this framework, deep networks are not robust to reinitializing with random weights while maintaining signs, and that weight distributions are highly non-independent even after only a few hundred iterations. Despite this, pre-training with blurred inputs or an auxiliary self-supervised task can approximate the changes in supervised networks, suggesting that these changes are label-agnostic, though labels significantly accelerate this process. Together, these results help to elucidate the network changes occurring during this pivotal initial period of learning.
****

Chordal-GCN: Exploiting sparsity in training large-scale graph convolutional networks
Xin Jiang*,Kewei Cheng*,Song Jiang*,Yizhou Sun
Despite the impressive success of graph convolutional networks (GCNs) on numerous applications, training on large-scale sparse networks remains challenging. Current algorithms require large memory space for storing GCN outputs as well as all the intermediate embeddings. Besides, most of these algorithms involves either random sampling or an approximation of the adjacency matrix, which might unfortunately lose important structure information. In this paper, we propose Chordal-GCN for semi-supervised node classification. The proposed model utilizes the exact graph structure (i.e., without sampling or approximation), while requires limited memory resources compared with the original GCN. Moreover, it leverages the sparsity pattern as well as the clustering structure of the graph. The proposed model first decomposes a large-scale sparse network into several small dense su

bgraphs (called cliques), and constructs a clique tree. By traversing the tree, GCN training is performed clique by clique, and connections between cliques are exploited via the tree hierarchy. Furthermore, we implement Chordal-GCN on large-scale datasets and demonstrate superior performance.

****

On The Difficulty of Warm-Starting Neural Network Training

Jordan T. Ash,Ryan P. Adams

In many real-world deployments of machine learning systems, data arrive piecemeal. These learning scenarios may be passive, where data arrive incrementally due to structural properties of the problem (e.g., daily financial data) or active, where samples are selected according to a measure of their quality (e.g., experimental design). In both of these cases, we are building a sequence of models that incorporate an increasing amount of data. We would like each of these models in the sequence to be performant and take advantage of all the data that are available to that point. Conventional intuition suggests that when solving a sequence of related optimization problems of this form, it should be possible to initialize using the solution of the previous iterate---to "warm start'' the optimization rather than initialize from scratch---and see reductions in wall-clock time. However, in practice this warm-starting seems to yield poorer generalization performance than models that have fresh random initializations, even though the final training losses are similar. While it appears that some hyperparameter settings allow a practitioner to close this generalization gap, they seem to only do so in regimes that damage the wall-clock gains of the warm start. Nevertheless, it is highly desirable to be able to warm-start neural network training, as it would dramatically reduce the resource usage associated with the construction of performant deep learning systems. In this work, we take a closer look at this empirical phenomenon and try to understand when and how it occurs. Although the present investigation did not lead to a solution, we hope that a thorough articulation of the problem will spur new research that may lead to improved methods that consume fewer resources during training.

****

NeuroFabric: Identifying Ideal Topologies for Training A Priori Sparse Networks

Mihailo Isakov,Michel A. Kinsy

Long training times of deep neural networks are a bottleneck in machine learning research. The major impediment to fast training is the quadratic growth of both memory and compute requirements of dense and convolutional layers with respect to their information bandwidth. Recently, training `a priori' sparse networks has been proposed as a method for allowing layers to retain high information bandwidth, while keeping memory and compute low. However, the choice of which sparse topology should be used in these networks is unclear. In this work, we provide a theoretical foundation for the choice of intra-layer topology. First, we derive a new sparse neural network initialization scheme that allows us to explore the space of very deep sparse networks. Next, we evaluate several topologies and show that seemingly similar topologies can often have a large difference in attainable accuracy. To explain these differences, we develop a data-free heuristic that can evaluate a topology independently from the dataset the network will be trained on. We then derive a set of requirements that make a good topology, and arrive at a single topology that satisfies all of them.

****

Distilled embedding: non-linear embedding factorization using knowledge distillation

Vasileios Lioutas,Ahmad Rashid,Krtin Kumar,Md Akmal Haidar,Mehdi Rezagholizadeh

Word-embeddings are a vital component of Natural Language Processing (NLP) systems and have been extensively researched. Better representations of words have come at the cost of huge memory footprints, which has made deploying NLP models on edge-devices challenging due to memory limitations. Compressing embedding matrices without sacrificing model performance is essential for successful commercial edge deployment. In this paper, we propose Distilled Embedding, an (input/output) embedding compression method based on low-rank matrix decomposition with an a

dded non-linearity. First, we initialize the weights of our decomposition by lea rning to reconstruct the full word-embedding and then fine-tune on the downstrea m task employing knowledge distillation on the factorized embedding. We conduct extensive experimentation with various compression rates on machine translation, using different data-sets with a shared word-embedding matrix for both embeddin g and vocabulary projection matrices. We show that the proposed technique outper forms conventional low-rank matrix factorization, and other recently proposed wo rd-embedding matrix compression methods.

****

RNNs Incrementally Evolving on an Equilibrium Manifold: A Panacea for Vanishing and Exploding Gradients?

Anil Kag,Ziming Zhang,Venkatesh Saligrama

Recurrent neural networks (RNNs) are particularly well-suited for modeling long- term dependencies in sequential data, but are notoriously hard to train because the error backpropagated in time either vanishes or explodes at an exponential r ate. While a number of works attempt to mitigate this effect through gated recur rent units, skip-connections, parametric constraints and design choices, we prop ose a novel incremental RNN (iRNN), where hidden state vectors keep track of inc remental changes, and as such approximate state-vector increments of Rosenblatt' s (1962) continuous-time RNNs. iRNN exhibits identity gradients and is able to a ccount for long-term dependencies (LTD). We show that our method is computationa lly efficient overcoming overheads of many existing methods that attempt to impr ove RNN training, while suffering no performance degradation. We demonstrate the utility of our approach with extensive experiments and show competitive perform ance against standard LSTMs on LTD and other non-LTD tasks.

****

Actor-Critic Approach for Temporal Predictive Clustering

Changhee Lee,Mihaela van der Schaar

Due to the wider availability of modern electronic health records (EHR), patient care data is often being stored in the form of time-series. Clustering such tim e-series data is crucial for patient phenotyping, anticipating patients’ prognos es by identifying “similar” patients, and designing treatment guidelines that ar e tailored to homogeneous patient subgroups. In this paper, we develop a deep le arning approach for clustering time-series data, where each cluster comprises pa tients who share similar future outcomes of interest (e.g., adverse events, the onset of comorbidities, etc.). The clustering is carried out by using our novel loss functions that encourage each cluster to have homogeneous future outcomes. We adopt actor-critic models to allow “back-propagation” through the sampling pr ocess that is required for assigning clusters to time-series inputs. Experiments on two real-world datasets show that our model achieves superior clustering per formance over state-of-the-art benchmarks and identifies meaningful clusters tha t can be translated into actionable information for clinical decision-making.

****

Adversarial Privacy Preservation under Attribute Inference Attack

Han Zhao,Jianfeng Chi,Yuan Tian,Geoffrey J. Gordon

With the prevalence of machine learning services, crowdsourced data containing s ensitive information poses substantial privacy challenges. Existing work focusin g on protecting against membership inference attacks under the rigorous framewor k of differential privacy are vulnerable to attribute inference attacks. In ligh t of the current gap between theory and practice, we develop a novel theoretical framework for privacy-preservation under the attack of attribute inference. Und er our framework, we propose a minimax optimization formulation to protect the g iven attribute and analyze its privacy guarantees against arbitrary adversaries. On the other hand, it is clear that privacy constraint may cripple utility when the protected attribute is correlated with the target variable. To this end, we also prove an information-theoretic lower bound to precisely characterize the f undamental trade-off between utility and privacy. Empirically, we extensively co nduct experiments to corroborate our privacy guarantee and validate the inherent

trade-offs in different privacy preservation algorithms. Our experimental results indicate that the adversarial representation learning approaches achieve the best trade-off in terms of privacy preservation and utility maximization.
****

Behavior-Guided Reinforcement Learning
Aldo Pacchiano,Jack Parker-Holder,Yunhao Tang,Anna Choromanska,Krzysztof Choromanski,Michael I. Jordan
We introduce a new approach for comparing reinforcement learning policies, using Wasserstein distances (WDs) in a newly defined latent behavioral space. We show that by utilizing the dual formulation of the WD, we can learn score functions over trajectories that can be in turn used to lead policy optimization towards (or away from) (un)desired behaviors. Combined with smoothed WDs, the dual formulation allows us to devise efficient algorithms that take stochastic gradient descent steps through WD regularizers. We incorporate these regularizers into two novel on-policy algorithms, Behavior-Guided Policy Gradient and Behavior-Guided Evolution Strategies, which we demonstrate can outperform existing methods in a variety of challenging environments. We also provide an open source demo.
****

Peer Loss Functions: Learning from Noisy Labels without Knowing Noise Rates
Yang Liu,Hongyi Guo
Learning with noisy labels is a common problem in supervised learning. Existing approaches require practitioners to specify noise rates, i.e., a set of parameters controlling the severity of label noises in the problem. In this work, we introduce a technique to learn from noisy labels that does not require a priori specification of the noise rates. In particular, we introduce a new family of loss functions that we name as peer loss functions. Our approach then uses a standard empirical risk minimization (ERM) framework with peer loss functions. Peer loss functions associate each training sample with a certain form of "peer" samples, which evaluate a classifier' predictions jointly. We show that, under mild conditions, performing ERM with peer loss functions on the noisy dataset leads to the optimal or a near optimal classifier as if performing ERM over the clean training data, which we do not have access to. To our best knowledge, this is the first result on "learning with noisy labels without knowing noise rates" with theoretical guarantees. We pair our results with an extensive set of experiments, where we compare with state-of-the-art techniques of learning with noisy labels. Our results show that peer loss functions based method consistently outperforms the baseline benchmarks. Peer loss provides a way to simplify model development when facing potentially noisy training labels, and can be promoted as a robust candidate loss function in such situations.
****

Learning to Contextually Aggregate Multi-Source Supervision for Sequence Labeling
Ouyu Lan*,Xiao Huang*,Bill Yuchen Lin,He Jiang,Xiang Ren
Sequence labeling is a fundamental framework for various natural language processing problems including part-of-speech tagging and named entity recognition. Its performance is largely influenced by the annotation quality and quantity in supervised learning scenarios. In many cases, ground truth labels are costly and time-consuming to collect or even non-existent, while imperfect ones could be easily accessed or transferred from different domains. A typical example is crowd-sourced datasets which have multiple annotations for each sentence which may be noisy or incomplete. Additionally, predictions from multiple source models in transfer learning can be seen as a case of multi-source supervision. In this paper, we propose a novel framework named Consensus Network (CONNET) to conduct training with imperfect annotations from multiple sources. It learns the representation for every weak supervision source and dynamically aggregates them by a context-aware attention mechanism. Finally, it leads to a model reflecting the consensus among multiple sources. We evaluate the proposed framework in two practical settings of multi-source learning: learning with crowd annotations and unsupervised cross-domain model adaptation. Extensive experimental results show that our model achieves significant improvements over existing methods in both se

ttings.
****
Extreme Tensoring for Low-Memory Preconditioning
Xinyi Chen,Naman Agarwal,Elad Hazan,Cyril Zhang,Yi Zhang
State-of-the-art models are now trained with billions of parameters, reaching hardware limits in terms of memory consumption. This has created a recent demand for memory-efficient optimizers. To this end, we investigate the limits and performance tradeoffs of memory-efficient adaptively preconditioned gradient methods. We propose \emph{extreme tensoring} for high-dimensional stochastic optimization, showing that an optimizer needs very little memory to benefit from adaptive preconditioning. Our technique applies to arbitrary models (not necessarily with tensor-shaped parameters), and is accompanied by regret and convergence guarantees, which shed light on the tradeoffs between preconditioner quality and expressivity. On a large-scale NLP model, we reduce the optimizer memory overhead by three orders of magnitude, without degrading performance.
****
Blockwise Adaptivity: Faster Training and Better Generalization in Deep Learning
Shuai Zheng,James T. Kwok
Stochastic methods with coordinate-wise adaptive stepsize (such as RMSprop and Adam) have been widely used in training deep neural networks. Despite their fast convergence, they can generalize worse than stochastic gradient descent. In this paper, by revisiting the design of Adagrad, we propose to split the network parameters into blocks, and use a blockwise
adaptive stepsize. Intuitively, blockwise adaptivity is less aggressive than adaptivity to individual coordinates, and can have a better balance between adaptivity and generalization. We show theoretically that the proposed blockwise adaptive gradient
descent has comparable regret in online convex learning and convergence rate for optimizing nonconvex objective as its counterpart with coordinate-wise adaptive stepsize, but is better up to some constant. We also study its uniform stability
and show that blockwise adaptivity can lead to lower generalization error than coordinate-wise adaptivity. Experimental results show that blockwise adaptive gradient descent converges faster and improves generalization performance over Nesterov's accelerated gradient and Adam.
****
Collapsed amortized variational inference for switching nonlinear dynamical systems
Zhe Dong,Bryan A. Seybold,Kevin P. Murphy,Hung H. Bui
We propose an efficient inference method for switching nonlinear dynamical systems. The key idea is to learn an inference network which can be used as a proposal distribution for the continuous latent variables, while performing exact marginalization of the discrete latent variables. This allows us to use the reparameterization trick, and apply end-to-end training with SGD. We show that this method can successfully segment time series data (including videos) into meaningful "regimes", due to the use of piece-wise nonlinear dynamics.
****
Non-Autoregressive Dialog State Tracking
Hung Le,Richard Socher,Steven C.H. Hoi
Recent efforts in Dialogue State Tracking (DST) for task-oriented dialogues have progressed toward open-vocabulary or generation-based approaches where the models can generate slot value candidates from the dialogue history itself. These approaches have shown good performance gain, especially in complicated dialogue domains with dynamic slot values. However, they fall short in two aspects: (1) they do not allow models to explicitly learn signals across domains and slots to detect potential dependencies among \textit{(domain, slot)} pairs; and (2) existing models follow auto-regressive approaches which incur high time cost when the dialogue evolves over multiple domains and multiple turns. In this paper, we propose a novel framework of Non-Autoregressive Dialog State Tracking (NADST) which

can factor in potential dependencies among domains and slots to optimize the mod
els towards better prediction of dialogue states as a complete set rather than s
eparate slots. In particular, the non-autoregressive nature of our method not on
ly enables decoding in parallel to significantly reduce the latency of DST for r
eal-time dialogue response generation, but also detect dependencies among slots
at token level in addition to slot and domain level. Our empirical results show
that our model achieves the state-of-the-art joint accuracy across all domains o
n the MultiWOZ 2.1 corpus, and the latency of our model is an order of magnitude
 lower than the previous state of the art as the dialogue history extends over t
ime.
****

Channel Equilibrium Networks
Wenqi Shao,Shitao Tang,Xingang Pan,Ping Tan,Xiaogang Wang,Ping Luo
Convolutional Neural Networks (CNNs) typically treat normalization methods such
as batch normalization (BN) and rectified linear function (ReLU) as building blo
cks. Previous work showed that this basic block would lead to channel-level spar
sity (i.e. channel of zero values), reducing computational complexity of CNNs. H
owever, over-sparse CNNs have many collapsed channels (i.e. many channels with u
ndesired zero values), impeding their learning ability. This problem is seldom e
xplored in the literature. To recover the collapsed channels and enhance learnin
g capacity, we propose a building block, Channel Equilibrium (CE), which takes
the output of a normalization layer as input and switches between two branches,
batch decorrelation (BD) branch and adaptive instance inverse (AII) branch. CE i
s able to prevent implicit channel-level sparsity in both experiments and theory
. It has several appealing properties. First, CE can be stacked after many norma
lization methods such as BN and Group Normalization (GN), and integrated into ma
ny advanced CNN architectures such as ResNet and MobileNet V2 to form a series o
f CE networks (CENets), consistently improving their performance. Second, extens
ive experiments show that CE achieves state-of-the-art results on various challe
nging benchmarks such as ImageNet and COCO. Third, we show an interesting connec
tion between CE and Nash Equilibrium, a well-known solution of a non-cooperative
 game. The models and code will be released soon.
****

Independence-aware Advantage Estimation
Pushi Zhang,Li Zhao,Guoqing Liu,Jiang Bian,Minglie Huang,Tao Qin,Tie-Yan Liu
Most of existing advantage function estimation methods in reinforcement learning
 suffer from the problem of high variance, which scales unfavorably with the tim
e horizon. To address this challenge, we propose to identify the independence pr
operty between current action and future states in environments, which can be fu
rther leveraged to effectively reduce the variance of the advantage estimation.
In particular, the recognized independence property can be naturally utilized to
 construct a novel importance sampling advantage estimator with close-to-zero va
riance even when the Monte-Carlo return signal yields a large variance. To furth
er remove the risk of the high variance introduced by the new estimator, we comb
ine it with existing Monte-Carlo estimator via a reward decomposition model lear
ned by minimizing the estimation variance. Experiments demonstrate that our meth
od achieves higher sample efficiency compared with existing advantage estimation
 methods in complex environments.
****

Bayesian Meta Sampling for Fast Uncertainty Adaptation
Zhenyi Wang,Yang Zhao,Ping Yu,Ruiyi Zhang,Changyou Chen
Meta learning has been making impressive progress for fast model adaptation. How
ever, limited work has been done on learning fast uncertainty adaption for Bayes
ian modeling. In this paper, we propose to achieve the goal by placing meta lear
ning on the space of probability measures, inducing the concept of meta sampling
 for fast uncertainty adaption. Specifically, we propose a Bayesian meta samplin
g framework consisting of two main components: a meta sampler and a sample adapt
er. The meta sampler is constructed by adopting a neural-inverse-autoregressive-
flow (NIAF) structure, a variant of the recently proposed neural autoregressive
flows, to efficiently generate meta samples to be adapted. The sample adapter mo

ves meta samples to task-specific samples, based on a newly proposed and general Bayesian sampling technique, called optimal-transport Bayesian sampling. The combination of the two components allows a simple learning procedure for the meta sampler to be developed, which can be efficiently optimized via standard back-propagation. Extensive experimental results demonstrate the efficiency and effectiveness of the proposed framework, obtaining better sample quality and faster uncertainty adaption compared to related methods.

****

Salient Explanation for Fine-grained Classification

Kanghan Oh,Sungchan Kim,Il-Seok Oh

Explaining the prediction of deep models has gained increasing attention to increase its applicability, even spreading it to life-affecting decisions. However there has been no attempt to pinpoint only the most discriminative features contributing specifically to separating different classes in a fine-grained classification task. This paper introduces a novel notion of salient explanation and proposes a simple yet effective salient explanation method called Gaussian light and shadow (GLAS), which estimates the spatial impact of deep models by the feature perturbation inspired by light and shadow in nature. GLAS provides a useful coarse-to-fine control benefiting from scalability of Gaussian mask. We also devised the ability to identify multiple instances through recursive GLAS. We prove the effectiveness of GLAS for fine-grained classification using the fine-grained classification dataset. To show the general applicability, we also illustrate that GLAS has state-of-the-art performance at high speed (about 0.5 sec per 224$\times$224 image) via the ImageNet Large Scale Visual Recognition Challenge.

****

Harnessing Structures for Value-Based Planning and Reinforcement Learning

Yuzhe Yang,Guo Zhang,Zhi Xu,Dina Katabi

Value-based methods constitute a fundamental methodology in planning and deep reinforcement learning (RL). In this paper, we propose to exploit the underlying structures of the state-action value function, i.e., Q function, for both planning and deep RL. In particular, if the underlying system dynamics lead to some global structures of the Q function, one should be capable of inferring the function better by leveraging such structures. Specifically, we investigate the low-rank structure, which widely exists for big data matrices. We verify empirically the existence of low-rank Q functions in the context of control and deep RL tasks. As our key contribution, by leveraging Matrix Estimation (ME) techniques, we propose a general framework to exploit the underlying low-rank structure in Q functions. This leads to a more efficient planning procedure for classical control, and additionally, a simple scheme that can be applied to value-based RL techniques to consistently achieve better performance on "low-rank" tasks. Extensive experiments on control tasks and Atari games confirm the efficacy of our approach.

****

The Dynamics of Signal Propagation in Gated Recurrent Neural Networks

Dar Gilboa,Bo Chang,Minmin Chen,Greg Yang,Samuel S. Schoenholz,Ed H. Chi,Jeffrey Pennington

Training recurrent neural networks (RNNs) on long sequence tasks is plagued with difficulties arising from the exponential explosion or vanishing of signals as they propagate forward or backward through the network. Many techniques have been proposed to ameliorate these issues, including various algorithmic and architectural modifications. Two of the most successful RNN architectures, the LSTM and the GRU, do exhibit modest improvements over vanilla RNN cells, but they still suffer from instabilities when trained on very long sequences. In this work, we develop a mean field theory of signal propagation in LSTMs and GRUs that enables us to calculate the time scales for signal propagation as well as the spectral properties of the state-to-state Jacobians. By optimizing these quantities in terms of the initialization hyperparameters, we derive a novel initialization scheme that eliminates or reduces training instabilities. We demonstrate the efficacy of our initialization scheme on multiple sequence tasks, on which it enables successful training while a standard initialization either fails completely or is

orders of magnitude slower. We also observe a beneficial effect on generalizati
on performance using this new initialization.
****
Economy Statistical Recurrent Units For Inferring Nonlinear Granger Causality
Saurabh Khanna,Vincent Y. F. Tan
Granger causality is a widely-used criterion for analyzing interactions in large
-scale networks. As most physical interactions are inherently nonlinear, we cons
ider the problem of inferring the existence of pairwise Granger causality betwee
n nonlinearly interacting stochastic processes from their time series measuremen
ts. Our proposed approach relies on modeling the embedded nonlinearities in the
measurements using a component-wise time series prediction model based on Statis
tical Recurrent Units (SRUs). We make a case that the network topology of Grange
r causal relations is directly inferrable from a structured sparse estimate of t
he internal parameters of the SRU networks trained to predict the processes' tim
e series measurements. We propose a variant of SRU, called economy-SRU, which, b
y design has considerably fewer trainable parameters, and therefore less prone t
o overfitting. The economy-SRU computes a low-dimensional sketch of its high-dim
ensional hidden state in the form of random projections to generate the feedback
for its recurrent processing. Additionally, the internal weight parameters of t
he economy-SRU are strategically regularized in a group-wise manner to facilitat
e the proposed network in extracting meaningful predictive features that are hig
hly time-localized to mimic real-world causal events. Extensive experiments are
carried out to demonstrate that the proposed economy-SRU based time series predi
ction model outperforms the MLP, LSTM and attention-gated CNN-based time series
models considered previously for inferring Granger causality.
****
Discriminability Distillation in Group Representation Learning
Manyuan Zhang■Guanglu Song■Yu Liu■Hang Zhou
Learning group representation is a commonly concerned issue in tasks where the b
asic unit is a group, set or sequence.
The computer vision community tries to tackle it by aggregating the elements in
a group based on an indicator either defined by human such as the quality or sal
iency of an element, or generated by a black box such as the attention score or
output of a RNN.

This article provides a more essential and explicable view.
We claim the most significant indicator to show whether the group representation
can be benefited from an element is not the quality, or an inexplicable score,
but the \textit{discrimiability}.
Our key insight is to explicitly design the \textit{discrimiability} using embed
ded class centroids on a proxy set,
and show the discrimiability distribution \textit{w.r.t.} the element space can
be distilled by a light-weight auxiliary distillation network.
This processing is called \textit{discriminability distillation learning} (DDL).
We show the proposed DDL can be flexibly plugged into many group based recogniti
on tasks without influencing the training procedure of the original tasks. Compr
ehensive experiments on set-to-set face recognition and action recognition valid
the advantage of DDL on both accuracy and efficiency, and it pushes forward the
state-of-the-art results on these tasks by an impressive margin.
****
Calibration, Entropy Rates, and Memory in Language Models
Mark Braverman,Xinyi Chen,Sham Kakade,Karthik Narasimhan,Cyril Zhang,Yi Zhang
Building accurate language models that capture meaningful long-term dependencies
is a core challenge in natural language processing. Towards this end, we presen
t a calibration-based approach to measure long-term discrepancies between a gene
rative sequence model and the true distribution, and use these discrepancies to
improve the model. Empirically, we show that state-of-the-art language models, i
ncluding LSTMs and Transformers, are \emph{miscalibrated}: the entropy rates of
their generations drift dramatically upward over time. We then provide provable
methods to mitigate this phenomenon. Furthermore, we show how this calibration-b

ased approach can also be used to measure the amount of memory that language mod
els use for prediction.
****
Efficient Saliency Maps for Explainable AI
T. Nathan Mundhenk,Barry Chen,Gerald Friedland
We describe an explainable AI saliency map method for use with deep convolutiona
l neural networks (CNN) that is much more efficient than popular gradient method
s. It is also quantitatively similar or better in accuracy. Our technique works
by measuring information at the end of each network scale. This is then combined
 into a single saliency map. We describe how saliency measures can be made more
efficient by exploiting Saliency Map Order Equivalence. Finally, we visualize i
ndividual scale/layer contributions by using a Layer Ordered Visualization of In
formation. This provides an interesting comparison of scale information contribu
tions within the network not provided by other saliency map methods. Our method
 is generally straight forward and should be applicable to the most commonly use
d CNNs. (Full source code is available at http://www.anonymous.submission.com).
****
Reinforcement Learning with Probabilistically Complete Exploration
Philippe Morere,Tom Blau,Gilad Francis,Fabio Ramos
Balancing exploration and exploitation remains a key challenge in reinforcement
learning (RL). State-of-the-art RL algorithms suffer from high sample complexity
, particularly in the sparse reward case, where they can do no better than to ex
plore in all directions until the first positive rewards are found. To mitigate
this, we propose Rapidly Randomly-exploring Reinforcement Learning (R3L). We for
mulate exploration as a search problem and leverage widely-used planning algorit
hms such as Rapidly-exploring Random Tree (RRT) to find initial solutions. These
 solutions are used as demonstrations to initialize a policy, then refined by a
generic RL algorithm, leading to faster and more stable convergence. We provide
theoretical guarantees of R3L exploration finding successful solutions, as well
as bounds for its sampling complexity. We experimentally demonstrate the method
outperforms classic and intrinsic exploration techniques, requiring only a fract
ion of exploration samples and achieving better asymptotic performance.
****
Unaligned Image-to-Sequence Transformation with Loop Consistency
Siyang Wang,Justin Lazarow,Kwonjoon Lee,Zhuowen Tu
We tackle the problem of modeling sequential visual phenomena. Given examples of
 a phenomena that can be divided into discrete time steps, we aim to take an inp
ut from any such time and realize this input at all other time steps in the sequ
ence. Furthermore, we aim to do this \textit{without} ground-truth aligned seque
nces --- avoiding the difficulties needed for gathering aligned data. This gener
alizes the unpaired image-to-image problem from generating pairs to generating s
equences. We extend cycle consistency to \textit{loop consistency} and alleviate
 difficulties associated with learning in the resulting long chains of computati
on. We show competitive results compared to existing image-to-image techniques w
hen modeling several different data sets including the Earth's seasons and aging
 of human faces.
****
Removing the Representation Error of GAN Image Priors Using the Deep Decoder
Max Daniels,Reinhard Heckel,Paul Hand
Generative models, such as GANs, have demonstrated impressive performance as nat
ural image priors for solving inverse problems such as image restoration and com
pressive sensing. Despite this performance, they can exhibit substantial represe
ntation error for both in-distribution and out-of-distribution images, because t
hey maintain explicit low-dimensional learned representations of a natural signa
l class. In this paper, we demonstrate a method for removing the representation
error of a GAN when used as a prior in inverse problems by modeling images as th
e linear combination of a GAN with a Deep Decoder. The deep decoder is an underp
arameterized and most importantly unlearned natural signal model similar to the
Deep Image Prior. No knowledge of the specific inverse problem is needed in the
 training of the GAN underlying our method. For compressive sensing and image s

uperresolution, our hybrid model exhibits consistently higher PSNRs than both th
e GAN priors and Deep Decoder separately, both on in-distribution and out-of-dis
tribution images. This model provides a method for extensibly and cheaply lever
aging both the benefits of learned and unlearned image recovery priors in invers
e problems.
****

MEMO: A Deep Network for Flexible Combination of Episodic Memories

Andrea Banino,Adrià Puigdomènech Badia,Raphael Köster,Martin J. Chadwick,Viniciu
s Zambaldi,Demis Hassabis,Caswell Barry,Matthew Botvinick,Dharshan Kumaran,Charl
es Blundell

Recent research developing neural network architectures with external memory hav
e often used the benchmark bAbI question and answering dataset which provides a
 challenging number of tasks requiring reasoning. Here we employed a classic asso
ciative inference task from the human neuroscience literature in order to more c
arefully probe the reasoning capacity of existing memory-augmented architectures
. This task is thought to capture the essence of reasoning -- the appreciation o
f distant relationships among elements distributed across multiple facts or memo
ries. Surprisingly, we found that current architectures struggle to reason over
 long distance associations. Similar results were obtained on a more complex task
 involving finding the shortest path between nodes in a path. We therefore devel
oped a novel architecture, MEMO, endowed with the capacity to reason over longer
 distances. This was accomplished with the addition of two novel components. Fir
st, it introduces a separation between memories/facts stored in external memory
 and the items that comprise these facts in external memory. Second, it makes use
 of an adaptive retrieval mechanism, allowing a variable number of ‘memory hops’
 before the answer is produced. MEMO is capable of solving our novel reasoning t
asks, as well as all 20 tasks in bAbI.
****

Superbloom: Bloom filter meets Transformer

John Anderson,Qingqing Huang,Walid Krichene,Steffen Rendle,Li Zhang

We extend the idea of word pieces in natural language models to machine learning
 tasks on opaque ids. This is achieved by applying hash functions to map each id
 to multiple hash tokens in a much smaller space, similarly to a Bloom filter. W
e show that by applying a multi-layer Transformer to these Bloom filter digests,
 we are able to obtain models with high accuracy. They outperform models of a si
milar size without hashing and, to a large degree, models of a much larger size
 trained using sampled softmax with the same computational budget. Our key observ
ation is that it is important to use a multi-layer Transformer for Bloom filter
 digests to remove ambiguity in the hashed input. We believe this provides an alt
ernative method to solving problems with large vocabulary size.
****

Longitudinal Enrichment of Imaging Biomarker Representations for Improved Alzhei
mer's Disease Diagnosis

Saad Elbeleidy,Lyujian Lu,L. Zoe Baker,Hua Wang,Feiping Nie

Longitudinal data is often available inconsistently across individuals resulting
 in ignoring of additionally available data. Alzheimer's Disease (AD) is a progr
essive disease that affects over 5 million patients in the US alone, and is the
 6th leading cause of death. Early detection of AD can significantly improve or e
xtend a patient's life so it is critical to use all available information about
 patients.

We propose an unsupervised method to learn a consistent representation by utiliz
ing inconsistent data through minimizing the ratio of p-Order Principal Compon
ents Analysis (PCA) and Locality Preserving Projections (LPP). Our method's repr
esentation can outperform the use of consistent data alone and does not require
 the use of complex tensor-specific approaches. We run experiments on patient dat
a from the Alzheimer’s Disease Neuroimaging Initiative (ADNI), which consists of
 inconsistent data, to predict patients' diagnosis.
****

Probabilistic Connection Importance Inference and Lossless Compression of Deep N
eural Networks

Xin Xing,Long Sha,Pengyu Hong,Zuofeng Shang,Jun S. Liu
Deep neural networks (DNNs) can be huge in size, requiring a considerable a moun t of energy and computational resources to operate, which limits their applicati ons in numerous scenarios. It is thus of interest to compress DNNs while maintai ning their performance levels. We here propose a probabilistic importance infer ence approach for pruning DNNs. Specifically, we test the significance of the re levance of a connection in a DNN to the DNN’s outputs using a nonparemtric scori ng testand keep only those significant ones. Experimental results show that the proposed approach achieves better lossless compression rates than existing techn iques
****

Generating Semantic Adversarial Examples with Differentiable Rendering
Lakshya Jain,Steven Chen,Wilson Wu,Uyeong Jang,Varun Chandrasekaran,Sanjit Seshi a,Somesh Jha
Machine learning (ML) algorithms, especially deep neural networks, have demonstr ated success in several domains. However, several types of attacks have raised c oncerns about deploying ML in safety-critical domains, such as autonomous drivin g and security. An attacker perturbs a data point slightly in the pixel space an d causes the ML algorithm to misclassify (e.g. a perturbed stop sign is classifi ed as a yield sign). These perturbed data points are called adversarial examples , and there are numerous algorithms in the literature for constructing adversari al examples and defending against them. In this paper we explore semantic advers arial examples (SAEs) where an attacker creates perturbations in the semantic sp ace. For example, an attacker can change the background of the image to be cloud ier to cause misclassification. We present an algorithm for constructing SAEs th at uses recent advances in differential rendering and inverse graphics.
****

Quantum algorithm for finding the negative curvature direction
Kaining Zhang,Min-Hsiu Hsieh,Liu Liu,Dacheng Tao
We present an efficient quantum algorithm aiming to find the negative curvature direction for escaping the saddle point, which is a critical subroutine for many second-order non-convex optimization algorithms. We prove that our algorithm co uld produce the target state corresponding to the negative curvature direction w ith query complexity O(polylog(d)ε^(-1)), where d is the dimension of the optimi zation function. The quantum negative curvature finding algorithm is exponential ly faster than any known classical method which takes time at least O(dε^(−1/2)) . Moreover, we propose an efficient algorithm to achieve the classical read-out of the target state. Our classical read-out algorithm runs exponentially faster on the degree of d than existing counterparts.
****

Dual-module Inference for Efficient Recurrent Neural Networks
Liu Liu,Lei Deng,Shuangchen Li,Jingwei Zhang,Yihua Yang,Zhenyu Gu,Yufei Ding,Yua n Xie
Using Recurrent Neural Networks (RNNs) in sequence modeling tasks is promising i n delivering high-quality results but challenging to meet stringent latency requ irements because of the memory-bound execution pattern of RNNs. We propose a big -little dual-module inference to dynamically skip unnecessary memory access and computation to speedup RNN inference. Leveraging the error-resilient feature of nonlinear activation functions used in RNNs, we propose to use a lightweight lit tle module that approximates the original RNN layer, which is referred to as the big module, to compute activations of the insensitive region that are more erro r-resilient. The expensive memory access and computation of the big module can b e reduced as the results are only used in the sensitive region. Our method can r educe the overall memory access by 40% on average and achieve 1.54x to 1.75x spe edup on CPU-based server platform with negligible impact on model quality.
****

GUIDEGAN: ATTENTION BASED SPATIAL GUIDANCE FOR IMAGE-TO-IMAGE TRANSLATION
Yu Lin,Yigong Wang,Yifan Li,Zhuoyi Wang,Yang Gao,Latifur Khan
Recently, Generative Adversarial Network (GAN) and numbers of its variants have been widely used to solve the image-to-image translation problem and achieved ex

traordinary results in both a supervised and unsupervised manner. However, most GAN-based methods suffer from the imbalance problem between the generator and discriminator in practice. Namely, the relative model capacities of the generator and discriminator do not match, leading to mode collapse and/or diminished gradients. To tackle this problem, we propose a GuideGAN based on attention mechanism. More specifically, we arm the discriminator with an attention mechanism so not only it estimates the probability that its input is real, but also does it create an attention map that highlights the critical features for such prediction. This attention map then assists the generator to produce more plausible and realistic images. We extensively evaluate the proposed GuideGAN framework on a number of image transfer tasks. Both qualitative results and quantitative comparison demonstrate the superiority of our proposed approach.
****
MixUp as Directional Adversarial Training
Guillaume Perrault-Archambault,Yongyi Mao,Hongyu Guo,Richong Zhang
MixUp is a data augmentation scheme in which pairs of training samples and their corresponding labels are mixed using linear coefficients. Without label mixing, MixUp becomes a more conventional scheme: input samples are moved but their original labels are retained. Because samples are preferentially moved in the direction of other classes \iffalse -- which are typically clustered in input space -- \fi we refer to this method as directional adversarial training, or DAT. We show that under two mild conditions, MixUp asymptotically convergences to a subset of DAT. We define untied MixUp (UMixUp), a superset of MixUp wherein training labels are mixed with different linear coefficients to those of their corresponding samples. We show that under the same mild conditions, untied MixUp converges to the entire class of DAT schemes. Motivated by the understanding that UMixUp is both a generalization of MixUp and a form of adversarial training, we experiment with different datasets and loss functions to show that UMixUp provides improved performance over MixUp. In short, we present a novel interpretation of MixUp as belonging to a class highly analogous to adversarial training, and on this basis we introduce a simple generalization which outperforms MixUp.
****
Towards Interpretable Molecular Graph Representation Learning
Emmanuel Noutahi,Dominique Beani,Julien Horwood,Prudencio Tossou
Recent work in graph neural networks (GNNs) has led to improvements in molecular activity and property prediction tasks. Unfortunately, GNNs often fail to capture the relative importance of interactions between molecular substructures, in part due to the absence of efficient intermediate pooling steps. To address these issues, we propose LaPool (Laplacian Pooling), a novel, data-driven, and interpretable hierarchical graph pooling method that takes into account both node features and graph structure to improve molecular understanding.
We benchmark LaPool and show that it not only outperforms recent GNNs on molecular graph understanding and prediction tasks but also remains highly competitive on other graph types. We then demonstrate the improved interpretability achieved with LaPool using both qualitative and quantitative assessments, highlighting its potential applications in drug discovery.
****
Representation Learning Through Latent Canonicalizations
Or Litany,Ari Morcos,Srinath Sridhar,Leonidas Guibas,Judy Hoffman
We seek to learn a representation on a large annotated data source that generalizes to a target domain using limited new supervision. Many prior approaches to this problem have focused on learning disentangled representations so that as individual factors vary in a new domain, only a portion of the representation need be updated. In this work, we seek the generalization power of disentangled representations, but relax the requirement of explicit latent disentanglement and instead encourage linearity of individual factors of variation by requiring them to be manipulable by learned linear transformations. We dub these transformations latent canonicalizers, as they aim to modify the value of a factor to a pre-determined (but arbitrary) canonical value (e.g., recoloring the image foreground to black). Assuming a source domain with access to meta-labels specifying the fact

ors of variation within an image, we demonstrate experimentally that our method helps reduce the number of observations needed to generalize to a similar target domain when compared to a number of supervised baselines.
****

Winning Privately: The Differentially Private Lottery Ticket Mechanism
Lovedeep Gondara,Ke Wang,Ricardo Silva Carvalho
We propose the differentially private lottery ticket mechanism (DPLTM). An end-to-end differentially private training paradigm based on the lottery ticket hypothesis. Using ``high-quality winners", selected via our custom score function, DPLTM significantly outperforms state-of-the-art. We show that DPLTM converges faster, allowing for early stopping with reduced privacy budget consumption. We further show that the tickets from DPLTM are transferable across datasets, domains, and architectures. Our extensive evaluation on several public datasets provides evidence to our claims.
****

Coherent Gradients: An Approach to Understanding Generalization in Gradient Descent-based Optimization
Satrajit Chatterjee
An open question in the Deep Learning community is why neural networks trained with Gradient Descent generalize well on real datasets even though they are capable of fitting random data. We propose an approach to answering this question based on a hypothesis about the dynamics of gradient descent that we call Coherent Gradients: Gradients from similar examples are similar and so the overall gradient is stronger in certain directions where these reinforce each other. Thus changes to the network parameters during training are biased towards those that (locally) simultaneously benefit many examples when such similarity exists. We support this hypothesis with heuristic arguments and perturbative experiments and outline how this can explain several common empirical observations about Deep Learning. Furthermore, our analysis is not just descriptive, but prescriptive. It suggests a natural modification to gradient descent that can greatly reduce overfitting.
****

Jelly Bean World: A Testbed for Never-Ending Learning
Emmanouil Antonios Platanios,Abulhair Saparov,Tom Mitchell
Machine learning has shown growing success in recent years. However, current machine learning systems are highly specialized, trained for particular problems or domains, and typically on a single narrow dataset. Human learning, on the other hand, is highly general and adaptable. Never-ending learning is a machine learning paradigm that aims to bridge this gap, with the goal of encouraging researchers to design machine learning systems that can learn to perform a wider variety of inter-related tasks in more complex environments. To date, there is no environment or testbed to facilitate the development and evaluation of never-ending learning systems. To this end, we propose the Jelly Bean World testbed. The Jelly Bean World allows experimentation over two-dimensional grid worlds which are filled with items and in which agents can navigate. This testbed provides environments that are sufficiently complex and where more generally intelligent algorithms ought to perform better than current state-of-the-art reinforcement learning approaches. It does so by producing non-stationary environments and facilitating experimentation with multi-task, multi-agent, multi-modal, and curriculum learning settings. We hope that this new freely-available software will prompt new research and interest in the development and evaluation of never-ending learning systems and more broadly, general intelligence systems.
****

Large-scale Pretraining for Neural Machine Translation with Tens of Billions of Sentence Pairs
Yuxian Meng,Xiangyuan Ren,Zijun Sun,Xiaoya Li,Arianna Yuan,Fei Wu,Jiwei Li
In this paper, we investigate the problem of training neural machine translation (NMT) systems with a dataset of more than 40 billion bilingual sentence pairs, which is larger than the largest dataset to date by orders of magnitude. Unprecedented challenges emerge in this situation compared to previous NMT work, includ

ing severe noise in the data and prohibitively long training time. We propose pr
actical solutions to handle these issues and demonstrate that large-scale pretra
ining significantly improves NMT performance. We are able to push the BLEU score
 of WMT17 Chinese-English dataset to 32.3, with a significant performance boos
t of +3.2 over existing state-of-the-art results.
****

Learning from Explanations with Neural Execution Tree
Ziqi Wang*,Yujia Qin*,Wenxuan Zhou,Jun Yan,Qinyuan Ye,Leonardo Neves,Zhiyuan Liu
,Xiang Ren
While deep neural networks have achieved impressive performance on a range of NL
P tasks, these data-hungry models heavily rely on labeled data, which restricts
 their applications in scenarios where data annotation is expensive. Natural lang
uage (NL) explanations have been demonstrated very useful additional supervision
, which can provide sufficient domain knowledge for generating more labeled data
 over new instances, while the annotation time only doubles. However, directly a
pplying them for augmenting model learning encounters two challenges: (1) NL exp
lanations are unstructured and inherently compositional, which asks for a modula
rized model to represent their semantics, (2) NL explanations often have large n
umbers of linguistic variants, resulting in low recall and limited generalizatio
n ability. In this paper, we propose a novel Neural Execution Tree (NExT) framew
ork to augment training data for text classification using NL explanations. Afte
r transforming NL explanations into executable logical forms by semantic parsing
, NExT generalizes different types of actions specified by the logical forms for
 labeling data instances, which substantially increases the coverage of each NL
explanation. Experiments on two NLP tasks (relation extraction and sentiment ana
lysis) demonstrate its superiority over baseline methods. Its extension to multi
-hop question answering achieves performance gain with light annotation effort.
****

A Coordinate-Free Construction of Scalable Natural Gradient
Kevin Luk,Roger Grosse
Most neural networks are trained using first-order optimization methods, which a
re sensitive to the parameterization of the model. Natural gradient descent is i
nvariant to smooth reparameterizations because it is defined in a coordinate-fre
e way, but tractable approximations are typically defined in terms of coordinate
 systems, and hence may lose the invariance properties. We analyze the invarianc
e properties of the Kronecker-Factored Approximate Curvature (K-FAC) algorithm b
y constructing the algorithm in a coordinate-free way. We explicitly construct a
 Riemannian metric under which the natural gradient matches the K-FAC update; in
variance to affine transformations of the activations follows immediately. We ex
tend our framework to analyze the invariance properties of K-FAC appied to convo
lutional networks and recurrent neural networks, as well as metrics other than t
he usual Fisher metric.
****

Discovering Motor Programs by Recomposing Demonstrations
Tanmay Shankar,Shubham Tulsiani,Lerrel Pinto,Abhinav Gupta
In this paper, we present an approach to learn recomposable motor primitives acr
oss large-scale and diverse manipulation demonstrations. Current approaches to d
ecomposing demonstrations into primitives often assume manually defined primitiv
es and bypass the difficulty of discovering these primitives. On the other hand,
 approaches in primitive discovery put restrictive assumptions on the complexity
 of a primitive, which limit applicability to narrow tasks. Our approach attempt
s to circumvent these challenges by jointly learning both the underlying motor p
rimitives and recomposing these primitives to form the original demonstration. T
hrough constraints on both the parsimony of primitive decomposition and the simp
licity of a given primitive, we are able to learn a diverse set of motor primiti
ves, as well as a coherent latent representation for these primitives. We demons
trate both qualitatively and quantitatively, that our learned primitives capture
 semantically meaningful aspects of a demonstration. This allows us to compose t
hese primitives in a hierarchical reinforcement learning setup to efficiently so
lve robotic manipulation tasks like reaching and pushing. Our results may be vie

wed at https://sites.google.com/view/discovering-motor-programs.
****
Adaptive Learned Bloom Filter (Ada-BF): Efficient Utilization of the Classifier
Zhenwei Dai,Anshumali Shrivastava
Recent work suggests improving the performance of Bloom filter by incorporating a machine learning model as a binary classifier. However, such learned Bloom filter does not take full advantage of the predicted probability scores. We proposed new algorithms that generalize the learned Bloom filter by using the complete spectrum of the scores regions. We proved our algorithms have lower False Positive Rate (FPR) and memory usage compared with the existing approaches to learned Bloom filter. We also demonstrated the improved performance of our algorithms on real-world datasets.
****
Convergence of Gradient Methods on Bilinear Zero-Sum Games
Guojun Zhang,Yaoliang Yu
Min-max formulations have attracted great attention in the ML community due to the rise of deep generative models and adversarial methods, while understanding the dynamics of gradient algorithms for solving such formulations has remained a grand challenge. As a first step, we restrict to bilinear zero-sum games and give a systematic analysis of popular gradient updates, for both simultaneous and alternating versions. We provide exact conditions for their convergence and find the optimal parameter setup and convergence rates. In particular, our results offer formal evidence that alternating updates converge "better" than simultaneous ones.
****
DSReg: Using Distant Supervision as a Regularizer
Yuxian Meng,Muyu Li,Xiaoya Li,Wei Wu,Fei Wu,Jiwei Li
In this paper, we aim at tackling a general issue in NLP tasks where some of the negative examples are highly similar to the positive examples, i.e., hard-negative examples). We propose the distant supervision as a regularizer (DSReg) approach to tackle this issue. We convert the original task to a multi-task learning problem, in which we first utilize the idea of distant supervision to retrieve hard-negative examples. The obtained hard-negative examples are then used as a regularizer, and we jointly optimize the original target objective of distinguishing positive examples from negative examples along with the auxiliary task objective of distinguishing soften positive examples (comprised of positive examples and hard-negative examples) from easy-negative examples. In the neural context, this can be done by feeding the final token representations to different output layers. Using this unbelievably simple strategy, we improve the performance of a range of different NLP tasks, including text classification, sequence labeling and reading comprehension.
****
Iterative Target Augmentation for Effective Conditional Generation
Kevin Yang,Wengong Jin,Kyle Swanson,Regina Barzilay,Tommi Jaakkola
Many challenging prediction problems, from molecular optimization to program synthesis, involve creating complex structured objects as outputs. However, available training data may not be sufficient for a generative model to learn all possible complex transformations. By leveraging the idea that evaluation is easier than generation, we show how a simple, broadly applicable, iterative target augmentation scheme can be surprisingly effective in guiding the training and use of such models. Our scheme views the generative model as a prior distribution, and employs a separately trained filter as the likelihood. In each augmentation step, we filter the model's outputs to obtain additional prediction targets for the next training epoch. Our method is applicable in the supervised as well as semi-supervised settings. We demonstrate that our approach yields significant gains over strong baselines both in molecular optimization and program synthesis. In particular, our augmented model outperforms the previous state-of-the-art in molecular optimization by over 10% in absolute gain.
****
Composing Task-Agnostic Policies with Deep Reinforcement Learning

Ahmed H. Qureshi,Jacob J. Johnson,Yuzhe Qin,Taylor Henderson,Byron Boots,Michael C. Yip
The composition of elementary behaviors to solve challenging transfer learning problems is one of the key elements in building intelligent machines. To date, there has been plenty of work on learning task-specific policies or skills but almost no focus on composing necessary, task-agnostic skills to find a solution to new problems. In this paper, we propose a novel deep reinforcement learning-based skill transfer and composition method that takes the agent's primitive policies to solve unseen tasks. We evaluate our method in difficult cases where training policy through standard reinforcement learning (RL) or even hierarchical RL is either not feasible or exhibits high sample complexity. We show that our method not only transfers skills to new problem settings but also solves the challenging environments requiring both task planning and motion control with high data efficiency.
****
The Local Elasticity of Neural Networks
Hangfeng He,Weijie Su
This paper presents a phenomenon in neural networks that we refer to as local elasticity. Roughly speaking, a classifier is said to be locally elastic if its prediction at a feature vector x' is not significantly perturbed, after the classifier is updated via stochastic gradient descent at a (labeled) feature vector x that is dissimilar to x' in a certain sense. This phenomenon is shown to persist for neural networks with nonlinear activation functions through extensive simulations on real-life and synthetic datasets, whereas this is not observed in linear classifiers. In addition, we offer a geometric interpretation of local elasticity using the neural tangent kernel (Jacot et al., 2018). Building on top of local elasticity, we obtain pairwise similarity measures between feature vectors, which can be used for clustering in conjunction with K-means. The effectiveness of the clustering algorithm on the MNIST and CIFAR-10 datasets in turn corroborates the hypothesis of local elasticity of neural networks on real-life data. Finally, we discuss some implications of local elasticity to shed light on several intriguing aspects of deep neural networks.
****
Gradient-Based Neural DAG Learning
Sébastien Lachapelle,Philippe Brouillard,Tristan Deleu,Simon Lacoste-Julien
We propose a novel score-based approach to learning a directed acyclic graph (DAG) from observational data. We adapt a recently proposed continuous constrained optimization formulation to allow for nonlinear relationships between variables using neural networks. This extension allows to model complex interactions while avoiding the combinatorial nature of the problem. In addition to comparing our method to existing continuous optimization methods, we provide missing empirical comparisons to nonlinear greedy search methods. On both synthetic and real-world data sets, this new method outperforms current continuous methods on most tasks while being competitive with existing greedy search methods on important metrics for causal inference.
****
On Concept-Based Explanations in Deep Neural Networks
Chih-Kuan Yeh,Been Kim,Sercan Arik,Chun-Liang Li,Pradeep Ravikumar,Tomas Pfister
Deep neural networks (DNNs) build high-level intelligence on low-level raw features. Understanding of this high-level intelligence can be enabled by deciphering the concepts they base their decisions on, as human-level thinking. In this paper, we study concept-based explainability for DNNs in a systematic framework. First, we define the notion of completeness, which quantifies how sufficient a particular set of concepts is in explaining a model's prediction behavior. Based on performance and variability motivations, we propose two definitions to quantify completeness. We show that under degenerate conditions, our method is equivalent to Principal Component Analysis. Next, we propose a concept discovery method that considers two additional constraints to encourage the interpretability of the discovered concepts. We use game-theoretic notions to aggregate over sets to define an importance score for each discovered concept, which we call \emph{Conce

ptSHAP}. On specifically-designed synthetic datasets and real-world text and ima
ge datasets, we validate the effectiveness of our framework in finding concepts
that are complete in explaining the decision, and interpretable.
****

Policy Message Passing: A New Algorithm for Probabilistic Graph Inference
Zhiwei Deng,Greg Mori
A general graph-structured neural network architecture operates on graphs throug
h two core components: (1) complex enough message functions; (2) a fixed informa
tion aggregation process. In this paper, we present the Policy Message Passing a
lgorithm, which takes a probabilistic perspective and reformulates the whole inf
ormation aggregation as stochastic sequential processes. The algorithm works on
a much larger search space, utilizes reasoning history to perform inference, and
 is robust to noisy edges. We apply our algorithm to multiple complex graph reas
oning and prediction tasks and show that our algorithm consistently outperforms
state-of-the-art graph-structured models by a significant margin.
****

Learning to Control Latent Representations for Few-Shot Learning of Named Entiti
es
Omar U. Florez,Erik Mueller
Humans excel in continuously learning with small data without forgetting how to
solve old problems.
However, neural networks require large datasets to compute latent representatio
ns across different tasks while minimizing a loss function. For example, a natur
al language understanding (NLU) system will often deal with emerging entities du
ring its deployment as interactions with users in realistic scenarios will gener
ate new and infrequent names, events, and locations. Here, we address this scena
rio by introducing a RL trainable controller that disentangles the representati
on learning of a neural encoder from its memory management role.

Our proposed solution is straightforward and simple: we train a controller to ex
ecute an optimal sequence of read and write operations on an external memory wit
h the goal of leveraging diverse activations from the past and provide accurat
e predictions. Our approach is named Learning to Control (LTC) and allows few-sh
ot learning with two degrees of memory plasticity. We experimentally show that
our system obtains accurate results for few-shot learning of entity recognition
in the Stanford Task-Oriented Dialogue dataset.
****

Amortized Nesterov's Momentum: Robust and Lightweight Momentum for Deep Learnin
g
Kaiwen Zhou,Yanghua Jin,Qinghua Ding,James Cheng
Stochastic Gradient Descent (SGD) with Nesterov's momentum is a widely used opti
mizer in deep learning, which is observed to have excellent generalization perfo
rmance. However, due to the large stochasticity, SGD with Nesterov's momentum is
 not robust, i.e., its performance may deviate significantly from the expectatio
n. In this work, we propose Amortized Nesterov's Momentum, a special variant of
Nesterov's momentum which has more robust iterates, faster convergence in the ea
rly stage and higher efficiency. Our experimental results show that this new mom
entum achieves similar (sometimes better) generalization performance with little
-to-no tuning. In the convex case, we provide optimal convergence rates for our
new methods and discuss how the theorems explain the empirical results.
****

Recurrent Event Network : Global Structure Inference Over Temporal Knowledge Gra
ph
Woojeong Jin,He Jiang,Meng Qu,Tong Chen,Changlin Zhang,Pedro Szekely,Xiang Ren
Modeling dynamically-evolving, multi-relational graph data has received a surge
of interests with the rapid growth of heterogeneous event data. However, predict
ing future events on such data requires global structure inference over time and
 the ability to integrate temporal and structural information, which are not yet
 well understood. We present Recurrent Event Network (RE-Net), a novel autoregre
ssive architecture for modeling temporal sequences of multi-relational graphs (e

.g., temporal knowledge graph), which can perform sequential, global structure i
nference over future time stamps to predict new events. RE-Net employs a recurre
nt event encoder to model the temporally conditioned joint probability distribut
ion for the event sequences, and equips the event encoder with a neighborhood ag
gregator for modeling the concurrent events within a time window associated with
 each entity. We apply teacher forcing for model training over historical data,
and infer graph sequences over future time stamps by sampling from the learned j
oint distribution in a sequential manner. We evaluate the proposed method via te
mporal link prediction on ■ve public datasets. Extensive experiments demonstrate
 the strength of RE-Net, especially on multi-step inference over future time sta
mps.
****

Composition-based Multi-Relational Graph Convolutional Networks
Shikhar Vashishth,Soumya Sanyal,Vikram Nitin,Partha Talukdar
Graph Convolutional Networks (GCNs) have recently been shown to be quite success
ful in modeling graph-structured data. However, the primary focus has been on ha
ndling simple undirected graphs. Multi-relational graphs are a more general and
prevalent form of graphs where each edge has a label and direction associated wi
th it. Most of the existing approaches to handle such graphs suffer from over-pa
rameterization and are restricted to learning representations of nodes only. In
this paper, we propose CompGCN, a novel Graph Convolutional framework which join
tly embeds both nodes and relations in a relational graph. CompGCN leverages a v
ariety of entity-relation composition operations from Knowledge Graph Embedding
techniques and scales with the number of relations. It also generalizes several
of the existing multi-relational GCN methods. We evaluate our proposed method on
 multiple tasks such as node classification, link prediction, and graph classifi
cation, and achieve demonstrably superior results. We make the source code of Co
mpGCN available to foster reproducible research.
****

Capsules with Inverted Dot-Product Attention Routing
Yao-Hung Hubert Tsai,Nitish Srivastava,Hanlin Goh,Ruslan Salakhutdinov
We introduce a new routing algorithm for capsule networks, in which a child caps
ule is routed to a parent based only on agreement between the parent's state and
 the child's vote.
The new mechanism 1) designs routing via inverted dot-product attention; 2) impo
ses Layer Normalization as normalization; and 3) replaces sequential iterative r
outing with concurrent iterative routing.
When compared to previously proposed routing algorithms, our method improves per
formance on benchmark datasets such as CIFAR-10 and CIFAR-100, and it performs a
t-par with a powerful CNN (ResNet-18) with 4x fewer parameters.
On a different task of recognizing digits from overlayed digit images, the propo
sed capsule model performs favorably against CNNs given the same number of layer
s and neurons per layer. We believe that our work raises the possibility of app
lying capsule networks to complex real-world tasks.
****

The Discriminative Jackknife: Quantifying Uncertainty in Deep Learning via Highe
r-Order Influence Functions
Ahmed M. Alaa,Mihaela van der Schaar
Deep learning models achieve high predictive accuracy in a broad spectrum of tas
ks, but rigorously quantifying their predictive uncertainty remains challenging.
 Usable estimates of predictive uncertainty should (1) cover the true prediction
 target with a high probability, and (2) discriminate between high- and low-conf
idence prediction instances. State-of-the-art methods for uncertainty quantifica
tion are based predominantly on Bayesian neural networks. However, Bayesian meth
ods may fall short of (1) and (2) — i.e., Bayesian credible intervals do not gua
rantee frequentist coverage, and approximate posterior inference may undermine d
iscriminative accuracy. To this end, this paper tackles the following question:
can we devise an alternative frequentist approach for uncertainty quantification
 that satisfies (1) and (2)?

To address this question, we develop the discriminative jackknife (DJ), a formal inference procedure that constructs predictive confidence intervals for a wide range of deep learning models, is easy to implement, and provides rigorous theoretical guarantees on (1) and (2). The DJ procedure uses higher-order influence functions (HOIFs) of the trained model parameters to construct a jackknife (leave-one-out) estimator of predictive confidence intervals. DJ computes HOIFs using a recursive formula that requires only oracle access to loss gradients and Hessian-vector products, hence it can be applied in a post-hoc fashion without compromising model accuracy or interfering with model training. Experiments demonstrate that DJ performs competitively compared to existing Bayesian and non-Bayesian baselines.
****

Insights on Visual Representations for Embodied Navigation Tasks

Erik Wijmans,Julian Straub,Irfan Essa,Dhruv Batra,Judy Hoffman,Ari Morcos

Recent advances in deep reinforcement learning require a large amount of training data and generally result in representations that are often over specialized to the target task. In this work, we study the underlying potential causes for this specialization by measuring the similarity between representations trained on related, but distinct tasks. We use the recently proposed projection weighted Canonical Correlation Analysis (PWCCA) to examine the task dependence of visual representations learned across different embodied navigation tasks. Surprisingly, we find that slight differences in task have no measurable effect on the visual representation for both SqueezeNet and ResNet architectures. We then empirically demonstrate that visual representations learned on one task can be effectively transferred to a different task. Interestingly, we show that if the tasks constrain the agent to spatially disjoint parts of the environment, differences in representation emerge for SqueezeNet models but less-so for ResNets, suggesting that ResNets feature inductive biases which encourage more task-agnostic representations, even in the context of spatially separated tasks. We generalize our analysis to examine permutations of an environment and find, surprisingly, permutations of an environment also do not influence the visual representation. Our analysis provides insight on the overfitting of representations in RL and provides suggestions of how to design tasks that induce task-agnostic representations.
****

Unsupervised Disentanglement of Pose, Appearance and Background from Images and Videos

Aysegul Dundar,Kevin J Shih,Animesh Garg,Robert Pottorf,Andrew Tao,Bryan Catanzaro

Unsupervised landmark learning is the task of learning semantic keypoint-like representations without the use of expensive keypoint-level annotations. A popular approach is to factorize an image into a pose and appearance data stream, then to reconstruct the image from the factorized components. The pose representation should capture a set of consistent and tightly localized landmarks in order to facilitate reconstruction of the input image. Ultimately, we wish for our learned landmarks to focus on the foreground object of interest. However, the reconstruction task of the entire image forces the model to allocate landmarks to model the background. This work explores the effects of factorizing the reconstruction task into separate foreground and background reconstructions, conditioning only the foreground reconstruction on the unsupervised landmarks. Our experiments demonstrate that the proposed factorization results in landmarks that are focused on the foreground object of interest. Furthermore, the rendered background quality is also improved, as the background rendering pipeline no longer requires the ill-suited landmarks to model its pose and appearance. We demonstrate this improvement in the context of the video-prediction.
****

On the Unintended Social Bias of Training Language Generation Models with News Articles

Omar U. Florez

There are concerns that neural language models may preserve some of the stereotypes of the underlying societies that generate the large corpora needed to train

these models. For example, gender bias is a significant problem when generating text, and its unintended memorization could impact the user experience of many applications (e.g., the smart-compose feature in Gmail).

In this paper, we introduce a novel architecture that decouples the representation learning of a neural model from its memory management role. This architecture allows us to update a memory module with an equal ratio across gender types addressing biased correlations directly in the latent space. We experimentally show that our approach can mitigate the gender bias amplification in the automatic generation of articles news while providing similar perplexity values when extending the Sequence2Sequence architecture.
****
Role-Wise Data Augmentation for Knowledge Distillation
Jie Fu,Xue Geng,Bohan Zhuang,Xingdi Yuan,Adam Trischler,Jie Lin,Vijay Chandrasekhar,Chris Pal
Knowledge Distillation (KD) is a common method for transferring the ``knowledge'' learned by one machine learning model (the teacher) into another model (the student), where typically, the teacher has a greater capacity (e.g., more parameters or higher bit-widths). To our knowledge, existing methods overlook the fact that although the student absorbs extra knowledge from the teacher, both models share the same input data -- and this data is the only medium by which the teacher's knowledge can be demonstrated. Due to the difference in model capacities, the student may not benefit fully from the same data points on which the teacher is trained. On the other hand, a human teacher may demonstrate a piece of knowledge with individualized examples adapted to a particular student, for instance, in terms of her cultural background and interests. Inspired by this behavior, we design data augmentation agents with distinct roles to facilitate knowledge distillation. Our data augmentation agents generate distinct training data for the teacher and student, respectively. We focus specifically on KD when the teacher network has greater precision (bit-width) than the student network.

We find empirically that specially tailored data points enable the teacher's knowledge to be demonstrated more effectively to the student. We compare our approach with existing KD methods on training popular neural architectures and demonstrate that role-wise data augmentation improves the effectiveness of KD over strong prior approaches. The code for reproducing our results will be made publicly available.
****
Attention Forcing for Sequence-to-sequence Model Training
Qingyun Dou,Yiting Lu,Joshua Efiong,Mark J.F. Gales
Auto-regressive sequence-to-sequence models with attention mechanism have achieved state-of-the-art performance in many tasks such as machine translation and speech synthesis. These models can be difficult to train. The standard approach, teacher forcing, guides a model with reference output history during training. The problem is that the model is unlikely to recover from its mistakes during inference, where the reference output is replaced by generated output. Several approaches deal with this problem, largely by guiding the model with generated output history. To make training stable, these approaches often require a heuristic schedule or an auxiliary classifier. This paper introduces attention forcing, which guides the model with generated output history and reference attention. This approach can train the model to recover from its mistakes, in a stable fashion, without the need for a schedule or a classifier. In addition, it allows the model to generate output sequences aligned with the references, which can be important for cascaded systems like many speech synthesis systems. Experiments on speech synthesis show that attention forcing yields significant performance gain. Experiments on machine translation show that for tasks where various re-orderings of the output are valid, guiding the model with generated output history is challenging, while guiding the model with reference attention is beneficial.
****
Topic Models with Survival Supervision: Archetypal Analysis and Neural Approache

s
George H. Chen,Linhong Li,Ren Zuo,Amanda Coston,Jeremy C. Weiss
We introduce two approaches to topic modeling supervised by survival analysis. B
oth approaches predict time-to-event outcomes while simultaneously learning topi
cs over features that help prediction. The high-level idea is to represent each
data point as a distribution over topics using some underlying topic model. Then
 each data point's distribution over topics is fed as input to a survival model.
 The topic and survival models are jointly learned. The two approaches we propos
e differ in the generality of topic models they can learn. The first approach fi
nds topics via archetypal analysis, a nonnegative matrix factorization method th
at optimizes over a wide class of topic models encompassing latent Dirichlet all
ocation (LDA), correlated topic models, and topic models based on the ``anchor w
ord'' assumption; the resulting survival-supervised variant solves an alternatin
g minimization problem. Our second approach builds on recent work that approxima
tes LDA in a neural net framework. We add a survival loss layer to this neural n
et to form an approximation to survival-supervised LDA. Both of our approaches c
an be combined with a variety of survival models. We demonstrate our approach on
 two survival datasets, showing that survival-supervised topic models can achiev
e competitive time-to-event prediction accuracy while outputting clinically inte
rpretable topics.
****
FSNet: Compression of Deep Convolutional Neural Networks by Filter Summary
Yingzhen Yang,Jiahui Yu,Nebojsa Jojic,Jun Huan,Thomas S. Huang
We present a novel method of compression of deep Convolutional Neural Networks (
CNNs) by weight sharing through a new representation of convolutional filters. T
he proposed method reduces the number of parameters of each convolutional layer
by learning a 1D vector termed Filter Summary (FS). The convolutional filters
are located in FS as overlapping 1D segments, and nearby filters in FS share w
eights in their overlapping regions in a natural way. The resultant neural netwo
rk based on such weight sharing scheme, termed Filter Summary CNNs or FSNet, has
 a FS in each convolution layer instead of a set of independent filters in the c
onventional convolution layer. FSNet has the same architecture as that of the ba
seline CNN to be compressed, and each convolution layer of FSNet has the same nu
mber of filters from FS as that of the basline CNN in the forward process. With
compelling computational acceleration ratio, the parameter space of FSNet is muc
h smaller than that of the baseline CNN. In addition, FSNet is quantization frie
ndly. FSNet with weight quantization leads to even higher compression ratio with
out noticeable performance loss. We further propose Differentiable FSNet where t
he way filters share weights is learned in a differentiable and end-to-end manne
r. Experiments demonstrate the effectiveness of FSNet in compression of CNNs for
 computer vision tasks including image classification and object detection, and
the effectiveness of DFSNet is evidenced by the task of Neural Architecture Sear
ch.
****
On the Need for Topology-Aware Generative Models for Manifold-Based Defenses
Uyeong Jang,Susmit Jha,Somesh Jha
ML algorithms or models, especially deep neural networks (DNNs), have shown sign
ificant promise in several areas. However, recently researchers have demonstrate
d that ML algorithms, especially DNNs, are vulnerable to adversarial examples (s
lightly perturbed samples that cause mis-classification). Existence of adversari
al examples has hindered deployment of ML algorithms in safety-critical sectors,
 such as security. Several defenses for adversarial examples exist in the litera
ture. One of the important classes of defenses are manifold-based defenses, wher
e a sample is "pulled back" into the data manifold before classifying. These def
enses rely on the manifold assumption (data lie in a manifold of lower dimension
 than the input space). These defenses use a generative model to approximate the
 input distribution. This paper asks the following question: do the generative m
odels used in manifold-based defenses need to be topology-aware? Our paper sugge
sts the answer is yes. We provide theoretical and empirical evidence to support
our claim.

****
Neural Execution of Graph Algorithms
Petar Veli■kovi■,Rex Ying,Matilde Padovano,Raia Hadsell,Charles Blundell
Graph Neural Networks (GNNs) are a powerful representational tool for solving pr
oblems on graph-structured inputs. In almost all cases so far, however, they hav
e been applied to directly recovering a final solution from raw inputs, without
explicit guidance on how to structure their problem-solving. Here, instead, we f
ocus on learning in the space of algorithms: we train several state-of-the-art G
NN architectures to imitate individual steps of classical graph algorithms, para
llel (breadth-first search, Bellman-Ford) as well as sequential (Prim's algorith
m). As graph algorithms usually rely on making discrete decisions within neighbo
urhoods, we hypothesise that maximisation-based message passing neural networks
are best-suited for such objectives, and validate this claim empirically. We als
o demonstrate how learning in the space of algorithms can yield new opportunitie
s for positive transfer between tasks---showing how learning a shortest-path alg
orithm can be substantially improved when simultaneously learning a reachability
 algorithm.
****
Objective Mismatch in Model-based Reinforcement Learning
Nathan Lambert,Brandon Amos,Omry Yadan,Roberto Calandra
Model-based reinforcement learning (MBRL) has been shown to be a powerful framew
ork for data-efficiently learning control of continuous tasks. Recent work in MB
RL has mostly focused on using more advanced function approximators and planning
 schemes, leaving the general framework virtually unchanged since its conception
. In this paper, we identify a fundamental issue of the standard MBRL framework
-- what we call the objective mismatch issue. Objective mismatch arises when one
 objective is optimized in the hope that a second, often uncorrelated, metric wi
ll also be optimized. In the context of MBRL, we characterize the objective mism
atch between training the forward dynamics model w.r.t. the likelihood of the on
e-step ahead prediction, and the overall goal of improving performance on a down
stream control task. For example, this issue can emerge with the realization tha
t dynamics models effective for a specific task do not necessarily need to be gl
obally accurate, and vice versa globally accurate models might not be sufficient
ly accurate locally to obtain good control performance on a specific task. In ou
r experiments, we study this objective mismatch issue and demonstrate that the l
ikelihood of the one-step ahead prediction is not always correlated with downstr
eam control performance. This observation highlights a critical flaw in the curr
ent MBRL framework which will require further research to be fully understood an
d addressed. We propose an initial method to mitigate the mismatch issue by re-w
eighting dynamics model training. Building on it, we conclude with a discussion
about other potential directions of future research for addressing this issue.
****
Molecular Graph Enhanced Transformer for Retrosynthesis Prediction
Kelong Mao,Peilin Zhao,Tingyang Xu,Yu Rong,Xi Xiao,Junzhou Huang
With massive possible synthetic routes in chemistry, retrosynthesis prediction i
s still a challenge for researchers. Recently, retrosynthesis prediction is form
ulated as a Machine Translation (MT) task. Namely, since each molecule can be re
presented as a Simpli■ed Molecular-Input Line-Entry System (SMILES) string, the
process of synthesis is analogized to a process of language translation from rea
ctants to products. However, the MT models that applied on SMILES data usually i
gnore the information of natural atomic connections and the topology of molecule
s. In this paper, we propose a Graph Enhanced Transformer (GET) framework, which
 adopts both the sequential and graphical information of molecules. Four differe
nt GET designs are proposed, which fuse the SMILES representations with atom emb
edding learned from our improved Graph Neural Network (GNN). Empirical results s
how that our model signi■cantly outperforms the Transformer model in test accura
cy.
****
Non-Sequential Melody Generation
Mitchell Billard,Robert Bishop,Moustafa Elsisy,Laura Graves,Antonina Kolokolova,

Vineel Nagisetty,Zachary Northcott,Heather Patey
In this paper we present a method for algorithmic melody generation using a gene rative adversarial network without recurrent components. Music generation has be en successfully done using recurrent neural networks, where the model learns seq uence information that can help create authentic sounding melodies. Here, we us e DCGAN architecture with dilated convolutions and towers to capture sequential information as spatial image information, and learn long-range dependencies in f ixed-length melody forms such as Irish traditional reel.
****

Compressing BERT: Studying the Effects of Weight Pruning on Transfer Learning
Mitchell A Gordon,Kevin Duh,Nicholas Andrews
Universal feature extractors, such as BERT for natural language processing and V GG for computer vision, have become effective methods for improving deep learnin g models without requiring more labeled data. A common paradigm is to pre-train a feature extractor on large amounts of data then fine-tune it as part of a deep learning model on some downstream task (i.e. transfer learning). While effectiv e, feature extractors like BERT may be prohibitively large for some deployment s cenarios. We explore weight pruning for BERT and ask: how does compression durin g pre-training affect transfer learning? We find that pruning affects transfer l earning in three broad regimes. Low levels of pruning (30-40%) do not affect pre -training loss or transfer to downstream tasks at all. Medium levels of pruning increase the pre-training loss and prevent useful pre-training information from being transferred to downstream tasks. High levels of pruning additionally preve nt models from fitting downstream datasets, leading to further degradation. Fina lly, we observe that fine-tuning BERT on a specific task does not improve its pr unability. We conclude that BERT can be pruned once during pre-training rather t han separately for each task without affecting performance.
****

Visual Explanation for Deep Metric Learning
Sijie Zhu,Taojiannan Yang,Chen Chen
This work explores the visual explanation for deep metric learning and its appli cations. As an important problem for learning representation, metric learning ha s attracted much attention recently, while the interpretation of such model is n ot as well studied as classification. To this end, we propose an intuitive idea to show where contributes the most to the overall similarity of two input images by decomposing the final activation. Instead of only providing the overall acti vation map of each image, we propose to generate point-to-point activation inten sity between two images so that the relationship between different regions is un covered. We show that the proposed framework can be directly deployed to a large range of metric learning applications and provides valuable information for und erstanding the model. Furthermore, our experiments show its effectiveness on two potential applications, i.e. cross-view pattern discovery and interactive retri eval.
****

Deep Innovation Protection
Sebastian Risi,Kenneth O. Stanley
Evolutionary-based optimization approaches have recently shown promising results in domains such as Atari and robot locomotion but less so in solving 3D tasks d irectly from pixels. This paper presents a method called Deep Innovation Protect ion (DIP) that allows training complex world models end-to-end for such 3D envi ronments. The main idea behind the approach is to employ multiobjective optimiza tion to temporally reduce the selection pressure on specific components in a wor ld model, allowing other components to adapt. We investigate the emergent repres entations of these evolved networks, which learn a model of the world without th e need for a specific forward-prediction loss.
****

Alternating Recurrent Dialog Model with Large-Scale Pre-Trained Language Models
Qingyang Wu,Yichi Zhang,Yu Li,Zhou Yu
Existing dialog system models require extensive human annotations and are diffic ult to generalize to different tasks. The recent success of large pre-trained la

nguage models such as BERT and GPT-2 have suggested the effectiveness of incorpo rating language priors in down-stream NLP tasks. However, how much pre-trained l anguage models can help dialog response generation is still under exploration. I n this paper, we propose a simple, general, and effective framework: Alternating Recurrent Dialog Model (ARDM). ARDM models each speaker separately and takes ad vantage of the large pre-trained language model. It requires no supervision from human annotations such as belief states or dialog acts to achieve effective con versations. ARDM outperforms or is on par with state-of-the-art methods on two p opular task-oriented dialog datasets: CamRest676 and MultiWOZ. Moreover, we can generalize ARDM to more challenging, non-collaborative tasks such as persuasion. In persuasion tasks, ARDM is capable of generating human-like responses to pers uade people to donate to a charity.
****

BERTScore: Evaluating Text Generation with BERT
Tianyi Zhang*,Varsha Kishore*,Felix Wu*,Kilian Q. Weinberger,Yoav Artzi
We propose BERTScore, an automatic evaluation metric for text generation. Analog ously to common metrics, BERTScore computes a similarity score for each token in the candidate sentence with each token in the reference sentence. However, inst ead of exact matches, we compute token similarity using contextual embeddings. W e evaluate using the outputs of 363 machine translation and image captioning sys tems. BERTScore correlates better with human judgments and provides stronger mod el selection performance than existing metrics. Finally, we use an adversarial p araphrase detection task and show that BERTScore is more robust to challenging e xamples compared to existing metrics.
****

Octave Graph Convolutional Network
Heng Chang,Yu Rong,Somayeh Sojoudi,Junzhou Huang,Wenwu Zhu
Many variants of Graph Convolutional Networks (GCNs) for representation learning have been proposed recently and have achieved fruitful results in various domai ns. Among them, spectral-based GCNs are constructed via convolution theorem upon theoretical foundation from the perspective of Graph Signal Processing (GSP). H owever, despite most of them implicitly act as low-pass filters that generate sm ooth representations for each node, there is limited development on the full usa ge of underlying information from low-frequency. Here, we first introduce the oc tave convolution on graphs in spectral domain. Accordingly, we present Octave Gr aph Convolutional Network (OctGCN), a novel architecture that learns representat ions for different frequency components regarding to weighted filters and graph wavelets bases. We empirically validate the importance of low-frequency componen ts in graph signals on semi-supervised node classification and demonstrate that our model achieves state-of-the-art performance in comparison with both spectral -based and spatial-based baselines.
****

Learning from Imperfect Annotations: An End-to-End Approach
Emmanouil Antonios Platanios,Maruan Al-Shedivat,Eric Xing,Tom Mitchell
Many machine learning systems today are trained on large amounts of human-annota ted data. Annotation tasks that require a high level of competency make data acq uisition expensive, while the resulting labels are often subjective, inconsiste nt, and may contain a variety of human biases. To improve data quality, practiti oners often need to collect multiple annotations per example and aggregate them before training models. Such a multi-stage approach results in redundant annotat ions and may often produce imperfect ``ground truth'' labels that limit the pote ntial of training supervised machine learning models. We propose a new end-to-en d framework that enables us to: (i) merge the aggregation step with model traini ng, thus allowing deep learning systems to learn to predict ground truth estimat es directly from the available data, and (ii) model difficulties of examples and learn representations of the annotators that allow us to estimate and take into account their competencies. Our approach is general and has many applications, including training more accurate models on crowdsourced data, ensemble learning, as well as classifier accuracy estimation from unlabeled data. We conduct an ex tensive experimental evaluation of our method on 5 crowdsourcing datasets of var

ied difficulty and show accuracy gains of up to 25% over the current state-of-th
e-art approaches for aggregating annotations, as well as significant reductions
in the required annotation redundancy.
****
Zeroth Order Optimization by a Mixture of Evolution Strategies
Jun-Kun Wang,Xiaoyun Li,Ping Li
Evolution strategies or zeroth-order optimization algorithms have become popular
 in some areas of optimization and machine learning where only the oracle of fun
ction value evaluations is available. The central idea in the design of the algo
rithms is by querying function values of some perturbed points in the neighborho
od of the current update and constructing a pseudo-gradient using the function v
alues. In recent years, there is a growing interest in developing new ways of pe
rturbation. Though the new perturbation methods are well motivating, most of the
m are criticized for lack of convergence guarantees even when the underlying fun
ction is convex. Perhaps the only methods that enjoy convergence guarantees are
the ones that sample the perturbed points uniformly from a unit sphere or from a
 multivariate Gaussian distribution with an isotropic covariance. In this work,
we tackle the non-convergence issue and propose sampling perturbed points from a
 mixture of distributions. Experiments show that our proposed method can identif
y the best perturbation scheme for the convergence and might also help to levera
ge the complementariness of different perturbation schemes.

****
Augmenting Non-Collaborative Dialog Systems with Explicit Semantic and Strategic
 Dialog History
Yiheng Zhou,Yulia Tsvetkov,Alan W Black,Zhou Yu
We study non-collaborative dialogs, where two agents have a conflict of interest
 but must strategically communicate to reach an agreement (e.g., negotiation). T
his setting poses new challenges for modeling dialog history because the dialog'
s outcome relies not only on the semantic intent, but also on tactics that conve
y the intent. We propose to model both semantic and tactic history using finite
 state transducers (FSTs). Unlike RNN, FSTs can explicitly represent dialog hist
ory through all the states traversed, facilitating interpretability of dialog st
ructure. We train FSTs on a set of strategies and tactics used in negotiation di
alogs. The trained FSTs show plausible tactic structure and can be generalized t
o other non-collaborative domains (e.g., persuasion). We evaluate the FSTs by in
corporating them in an automated negotiating system that attempts to sell produc
ts and a persuasion system that persuades people to donate to a charity. Experim
ents show that explicitly modeling both semantic and tactic history is an effect
ive way to improve both dialog policy planning and generation performance.
****
Machine Truth Serum
Tianyi Luo,Yang Liu
Wisdom of the crowd revealed a striking fact that the majority answer from a cro
wd is often more accurate than any individual expert. We observed the same story
 in machine learning - ensemble methods leverage this idea to combine multiple l
earning algorithms to obtain better classification performance. Among many popul
ar examples is the celebrated Random Forest, which applies the majority voting r
ule in aggregating different decision trees to make the final prediction. Noneth
eless, these aggregation rules would fail when the majority is more likely to be
 wrong. In this paper, we extend the idea proposed in Bayesian Truth Serum that
"a surprisingly more popular answer is more likely the true answer" to classific
ation problems. The challenge for us is to define or detect when an answer shoul
d be considered as being "surprising". We present two machine learning aided met
hods which aim to reveal the truth when it is minority instead of majority who h
as the true answer. Our experiments over real-world datasets show that better cl
assification performance can be obtained compared to always trusting the majorit
y voting. Our proposed methods also outperform popular ensemble algorithms. Our
approach can be generically applied as a subroutine in ensemble methods to repla
ce majority voting rule.

****
Prediction, Consistency, Curvature: Representation Learning for Locally-Linear Control
Nir Levine,Yinlam Chow,Rui Shu,Ang Li,Mohammad Ghavamzadeh,Hung Bui
Many real-world sequential decision-making problems can be formulated as optimal control with high-dimensional observations and unknown dynamics. A promising approach is to embed the high-dimensional observations into a lower-dimensional latent representation space, estimate the latent dynamics model, then utilize this model for control in the latent space. An important open question is how to learn a representation that is amenable to existing control algorithms? In this paper, we focus on learning representations for locally-linear control algorithms, such as iterative LQR (iLQR). By formulating and analyzing the representation learning problem from an optimal control perspective, we establish three underlying principles that the learned representation should comprise: 1) accurate prediction in the observation space, 2) consistency between latent and observation space dynamics, and 3) low curvature in the latent space transitions. These principles naturally correspond to a loss function that consists of three terms: prediction, consistency, and curvature (PCC). Crucially, to make PCC tractable, we derive an amortized variational bound for the PCC loss function. Extensive experiments on benchmark domains demonstrate that the new variational-PCC learning algorithm benefits from significantly more stable and reproducible training, and leads to superior control performance. Further ablation studies give support to the importance of all three PCC components for learning a good latent space for control.
****
GraphZoom: A Multi-level Spectral Approach for Accurate and Scalable Graph Embedding
Chenhui Deng,Zhiqiang Zhao,Yongyu Wang,Zhiru Zhang,Zhuo Feng
Graph embedding techniques have been increasingly deployed in a multitude of different applications that involve learning on non-Euclidean data. However, existing graph embedding models either fail to incorporate node attribute information during training or suffer from node attribute noise, which compromises the accuracy. Moreover, very few of them scale to large graphs due to their high computational complexity and memory usage. In this paper we propose GraphZoom, a multi-level framework for improving both accuracy and scalability of unsupervised graph embedding algorithms. GraphZoom first performs graph fusion to generate a new graph that effectively encodes the topology of the original graph and the node attribute information. This fused graph is then repeatedly coarsened into much smaller graphs by merging nodes with high spectral similarities. GraphZoom allows any existing embedding methods to be applied to the coarsened graph, before it progressively refine the embeddings obtained at the coarsest level to increasingly finer graphs. We have evaluated our approach on a number of popular graph datasets for both transductive and inductive tasks. Our experiments show that GraphZoom can substantially increase the classification accuracy and significantly accelerate the entire graph embedding process by up to $40.8 \times$, when compared to the state-of-the-art unsupervised embedding methods.
****
Sensible adversarial learning
Jungeum Kim,Xiao Wang
The trade-off between robustness and standard accuracy has been consistently reported in the machine learning literature. Although the problem has been widely studied to understand and explain this trade-off, no studies have shown the possibility of a no trade-off solution. In this paper, motivated by the fact that the high dimensional distribution is poorly represented by limited data samples, we introduce sensible adversarial learning and demonstrate the synergistic effect between pursuits of natural accuracy and robustness. Specifically, we define a sensible adversary which is useful for learning a defense model and keeping a high natural accuracy simultaneously. We theoretically establish that the Bayes rule is the most robust multi-class classifier with the 0-1 loss under sensible adversarial learning. We propose a novel and efficient algorithm that trains a robu

st model with sensible adversarial examples, without a significant drop in natur al accuracy. Our model on CIFAR10 yields state-of-the-art results against variou s attacks with perturbations restricted to l_∞ with $\varepsilon = 8/255$, e.g., the robust a ccuracy 65.17% against PGD attacks as well as the natural accuracy 91.51%.

****
Attention Interpretability Across NLP Tasks
Shikhar Vashishth,Shyam Upadhyay,Gaurav Singh Tomar,Manaal Faruqui
The attention layer in a neural network model provides insights into the model's reasoning behind its prediction, which are usually criticized for being opaque. Recently, seemingly contradictory viewpoints have emerged about the interpretab ility of attention weights (Jain & Wallace, 2019; Vig & Belinkov, 2019). Amid su ch confusion arises the need to understand attention mechanism more systematical ly. In this work, we attempt to fill this gap by giving a comprehensive explanat ion which justifies both kinds of observations (i.e., when is attention interpre table and when it is not). Through a series of experiments on diverse NLP tasks, we validate our observations and reinforce our claim of interpretability of att ention through manual evaluation.
****
MoET: Interpretable and Verifiable Reinforcement Learning via Mixture of Expert Trees
Marko Vasic,Andrija Petrovic,Kaiyuan Wang,Mladen Nikolic,Rishabh Singh,Sarfraz K hurshid
Deep Reinforcement Learning (DRL) has led to many recent breakthroughs on comple x control tasks, such as defeating the best human player in the game of Go. Howe ver, decisions made by the DRL agent are not explainable, hindering its applicab ility in safety-critical settings. Viper, a recently proposed technique, constru cts a decision tree policy by mimicking the DRL agent. Decision trees are interp retable as each action made can be traced back to the decision rule path that le ad to it. However, one global decision tree approximating the DRL policy has sig nificant limitations with respect to the geometry of decision boundaries. We pro pose MoET, a more expressive, yet still interpretable model based on Mixture of Experts, consisting of a gating function that partitions the state space, and mu ltiple decision tree experts that specialize on different partitions. We propose a training procedure to support non-differentiable decision tree experts and in tegrate it into imitation learning procedure of Viper. We evaluate our algorithm on four OpenAI gym environments, and show that the policy constructed in such a way is more performant and better mimics the DRL agent by lowering mispredictio ns and increasing the reward. We also show that MoET policies are amenable for v erification using off-the-shelf automated theorem provers such as Z3.
****
AdaScale SGD: A Scale-Invariant Algorithm for Distributed Training
Tyler B. Johnson,Pulkit Agrawal,Haijie Gu,Carlos Guestrin
When using distributed training to speed up stochastic gradient descent, learnin g rates must adapt to new scales in order to maintain training effectiveness. Re -tuning these parameters is resource intensive, while fixed scaling rules often degrade model quality. We propose AdaScale SGD, a practical and principled algor ithm that is approximately scale invariant. By continually adapting to the gradi ent's variance, AdaScale often trains at a wide range of scales with nearly iden tical results. We describe this invariance formally through AdaScale's convergen ce bounds. As the batch size increases, the bounds maintain final objective valu es, while smoothly transitioning away from linear speed-ups. In empirical compar isons, AdaScale trains well beyond the batch size limits of popular "linear lear ning rate scaling" rules. This includes large-scale training without model degra dation for machine translation, image classification, object detection, and spee ch recognition tasks. The algorithm introduces negligible computational overhead and no tuning parameters, making AdaScale an attractive choice for large-scale training.

****

INTERNAL-CONSISTENCY CONSTRAINTS FOR EMERGENT COMMUNICATION
Charles Lovering,Ellie Pavlick
When communicating, humans rely on internally-consistent language representation
s. That is, as speakers, we expect listeners to behave the same way we do when w
e listen. This work proposes several methods for encouraging such internal consi
stency in dialog agents in an emergent communication setting. We consider two hy
potheses about the effect of internal-consistency constraints: 1) that they impr
ove agents' ability to refer to unseen referents, and 2) that they improve agent
s' ability to generalize across communicative roles (e.g. performing as a speake
r de- spite only being trained as a listener). While we do not find evidence in
favor of the former, our results show significant support for the latter.
****

Bio-Inspired Hashing for Unsupervised Similarity Search
Chaitanya K. Ryali,John J. Hopfield,Dmitry Krotov
The fruit fly Drosophila's olfactory circuit has inspired a new locality sensiti
ve hashing (LSH) algorithm, FlyHash. In contrast with classical LSH algorithms t
hat produce low dimensional hash codes, FlyHash produces sparse high-dimensional
 hash codes and has also been shown to have superior empirical performance compa
red to classical LSH algorithms in similarity search. However, FlyHash uses rand
om projections and cannot learn from data. Building on inspiration from FlyHash
and the ubiquity of sparse expansive representations in neurobiology, our work p
roposes a novel hashing algorithm BioHash that produces sparse high dimensional
hash codes in a data-driven manner. We show that BioHash outperforms previously
 published benchmarks for various hashing methods. Since our learning algorithm
is based on a local and biologically plausible synaptic plasticity rule, our wor
k provides evidence for the proposal that LSH might be a computational reason fo
r the abundance of sparse expansive motifs in a variety of biological systems. W
e also propose a convolutional variant BioConvHash that further improves perform
ance. From the perspective of computer science, BioHash and BioConvHash are fas
t, scalable and yield compressed binary representations that are useful for simi
larity search.
****

Simplicial Complex Networks
Mohammad Firouzi,Sadra Boreiri,Hamed Firouzi
Universal approximation property of neural networks is one of the motivations to
 use these models in various real-world problems. However, this property is not
the only characteristic that makes neural networks unique as there is a wide ran
ge of other approaches with similar property. Another characteristic which makes
 these models interesting is that they can be trained with the backpropagation a
lgorithm which allows an efficient gradient computation and gives these universa
l approximators the ability to efficiently learn complex manifolds from a large
amount of data in different domains. Despite their abundant use in practice, neu
ral networks are still not well understood and a broad range of ongoing research
 is to study the interpretability of neural networks. On the other hand, topolog
ical data analysis (TDA) relies on strong theoretical framework of (algebraic) t
opology along with other mathematical tools for analyzing possibly complex datas
ets. In this work, we leverage a universal approximation theorem originating fro
m algebraic topology to build a connection between TDA and common neural network
 training framework. We introduce the notion of automatic subdivisioning and dev
ise a particular type of neural networks for regression tasks: Simplicial Comple
x Networks (SCNs). SCN's architecture is defined with a set of bias functions al
ong with a particular policy during the forward pass which alternates the common
 architecture search framework in neural networks. We believe the view of SCNs c
an be used as a step towards building interpretable deep learning models. Finall
y, we verify its performance on a set of regression problems.
****

BEYOND SUPERVISED LEARNING: RECOGNIZING UNSEEN ATTRIBUTE-OBJECT PAIRS WITH VISIO
N-LANGUAGE FUSION AND ATTRACTOR NETWORKS
Hui Chen,Zhixiong Nan,Nanning Zheng
This paper handles a challenging problem, unseen attribute-object pair recogniti

on, which asks a model to simultaneously recognize the attribute type and the object type of a given image while this attribute-object pair is not included in the training set. In the past years, the conventional classifier-based methods, which recognize unseen attribute-object pairs by composing separately-trained attribute classifiers and object classifiers, are strongly frustrated. Different from conventional methods, we propose a generative model with a visual pathway and a linguistic pathway. In each pathway, the attractor network is involved to learn the intrinsic feature representation to explore the inner relationship between the attribute and the object. With the learned features in both pathways, the unseen attribute-object pair is recognized by finding out the pair whose linguistic feature closely matches the visual feature of the given image. On two public datasets, our model achieves impressive experiment results, notably outperforming the state-of-the-art methods.

****

Underwhelming Generalization Improvements From Controlling Feature Attribution

Joseph D Viviano,Becks Simpson,Francis Dutil,Yoshua Bengio,Joseph Paul Cohen

Overfitting is a common issue in machine learning, which can arise when the model learns to predict class membership using convenient but spuriously-correlated image features instead of the true image features that denote a class. These are typically visualized using saliency maps. In some object classification tasks such as for medical images, one may have some images with masks, indicating a region of interest, i.e., which part of the image contains the most relevant information for the classification. We describe a simple method for taking advantage of such auxiliary labels, by training networks to ignore the distracting features which may be extracted outside of the region of interest, on the training images for which such masks are available. This mask information is only used during training and has an impact on generalization accuracy in a dataset-dependent way. We observe an underwhelming relationship between controlling saliency maps and improving generalization performance.

****

Graph Constrained Reinforcement Learning for Natural Language Action Spaces

Prithviraj Ammanabrolu,Matthew Hausknecht

Interactive Fiction games are text-based simulations in which an agent interacts with the world purely through natural language. They are ideal environments for studying how to extend reinforcement learning agents to meet the challenges of natural language understanding, partial observability, and action generation in combinatorially-large text-based action spaces. We present KG-A2C, an agent that builds a dynamic knowledge graph while exploring and generates actions using a template-based action space. We contend that the dual uses of the knowledge graph to reason about game state and to constrain natural language generation are the keys to scalable exploration of combinatorially large natural language actions. Results across a wide variety of IF games show that KG-A2C outperforms current IF agents despite the exponential increase in action space size.

****

Solving Packing Problems by Conditional Query Learning

Dongda Li,Changwei Ren,Zhaoquan Gu,Yuexuan Wang,Francis Lau

Neural Combinatorial Optimization (NCO) has shown the potential to solve traditional NP-hard problems recently. Previous studies have shown that NCO outperforms heuristic algorithms in many combinatorial optimization problems such as the routing problems. However, it is less efficient for more complicated problems such as packing, one type of optimization problem that faces mutual conditioned action space. In this paper, we propose a Conditional Query Learning (CQL) method to handle the packing problem for both 2D and 3D settings. By embedding previous actions as a conditional query to the attention model, we design a fully end-to-end model and train it for 2D and 3D packing via reinforcement learning respectively. Through extensive experiments, the results show that our method could achieve lower bin gap ratio and variance for both 2D and 3D packing. Our model improves 7.2% space utilization ratio compared with genetic algorithm for 3D packing (30 boxes case), and reduces more than 10% bin gap ratio in almost every case compared with extant learning approaches. In addition, our model shows great scala

bility to packing box number. Furthermore, we provide a general test environment of 2D and 3D packing for learning algorithms. All source code of the model and the test environment is released.
****

Task-Relevant Adversarial Imitation Learning
Konrad Zolna,Scott Reed,Alexander Novikov,Ziyu Wang,Sergio Gómez,David Budden,Serkan Cabi,Misha Denil,Nando de Freitas
We show that a critical problem in adversarial imitation from high-dimensional sensory data is the tendency of discriminator networks to distinguish agent and expert behaviour using task-irrelevant features beyond the control of the agent. We analyze this problem in detail and propose a solution as well as several baselines that outperform standard Generative Adversarial Imitation Learning (GAIL). Our proposed solution, Task-Relevant Adversarial Imitation Learning (TRAIL), uses a constrained optimization objective to overcome task-irrelevant features. Comprehensive experiments show that TRAIL can solve challenging manipulation tasks from pixels by imitating human operators, where other agents such as behaviour cloning (BC), standard GAIL, improved GAIL variants including our newly proposed baselines, and Deterministic Policy Gradients from Demonstrations (DPGfD) fail to find solutions, even when the other agents have access to task reward.
****

Generative Restricted Kernel Machines
Arun Pandey,Joachim Schreurs,Johan A.K. Suykens
We introduce a novel framework for generative models based on Restricted Kernel Machines (RKMs) with multi-view generation and uncorrelated feature learning capabilities, called Gen-RKM. To incorporate multi-view generation, this mechanism uses a shared representation of data from various views. The mechanism is flexible to incorporate both kernel-based, (deep) neural network and convolutional based models within the same setting. To update the parameters of the network, we propose a novel training procedure which jointly learns the features and shared representation. Experiments demonstrate the potential of the framework through qualitative evaluation of generated samples.
****

Towards Fast Adaptation of Neural Architectures with Meta Learning
Dongze Lian,Yin Zheng,Yintao Xu,Yanxiong Lu,Leyu Lin,Peilin Zhao,Junzhou Huang,Shenghua Gao
Recently, Neural Architecture Search (NAS) has been successfully applied to multiple artificial intelligence areas and shows better performance compared with hand-designed networks. However, the existing NAS methods only target a specific task. Most of them usually do well in searching an architecture for single task but are troublesome for multiple datasets or multiple tasks. Generally, the architecture for a new task is either searched from scratch, which is neither efficient nor flexible enough for practical application scenarios, or borrowed from the ones searched on other tasks, which might be not optimal. In order to tackle the transferability of NAS and conduct fast adaptation of neural architectures, we propose a novel Transferable Neural Architecture Search method based on meta-learning in this paper, which is termed as T-NAS. T-NAS learns a meta-architecture that is able to adapt to a new task quickly through a few gradient steps, which makes the transferred architecture suitable for the specific task. Extensive experiments show that T-NAS achieves state-of-the-art performance in few-shot learning and comparable performance in supervised learning but with 50x less searching cost, which demonstrates the effectiveness of our method.
****

A Functional Characterization of Randomly Initialized Gradient Descent in Deep ReLU Networks
Justin Sahs,Aneel Damaraju,Ryan Pyle,Onur Tavaslioglu,Josue Ortega Caro,Hao Yang Lu,Ankit Patel
Despite their popularity and successes, deep neural networks are poorly understood theoretically and treated as 'black box' systems. Using a functional view of these networks gives us a useful new lens with which to understand them. This allows us us to theoretically or experimentally probe properties of these networks

, including the effect of standard initializations, the value of depth, the unde
rlying loss surface, and the origins of generalization. One key result is that g
eneralization results from smoothness of the functional approximation, combined
with a flat initial approximation. This smoothness increases with number of unit
s, explaining why massively overparamaterized networks continue to generalize we
ll.
****

Variational Hetero-Encoder Randomized GANs for Joint Image-Text Modeling
Hao Zhang,Bo Chen,Long Tian,Zhengjue Wang,Mingyuan Zhou
For bidirectional joint image-text modeling, we develop variational hetero-encod
er (VHE) randomized generative adversarial network (GAN), a versatile deep gener
ative model that integrates a probabilistic text decoder, probabilistic image en
coder, and GAN into a coherent end-to-end multi-modality learning framework. VHE
 randomized GAN (VHE-GAN) encodes an image to decode its associated text, and fe
eds the variational posterior as the source of randomness into the GAN image gen
erator. We plug three off-the-shelf modules, including a deep topic model, a lad
der-structured image encoder, and StackGAN++, into VHE-GAN, which already achiev
es competitive performance. This further motivates the development of VHE-raster
-scan-GAN that generates photo-realistic images in not only a multi-scale low-to
-high-resolution manner, but also a hierarchical-semantic coarse-to-fine fashion
. By capturing and relating hierarchical semantic and visual concepts with end-t
o-end training, VHE-raster-scan-GAN achieves state-of-the-art performance in a w
ide variety of image-text multi-modality learning and generation tasks.
****

Toward Understanding Generalization of Over-parameterized Deep ReLU network trai
ned with SGD in Student-teacher Setting
Yuandong Tian
To analyze deep ReLU network, we adopt a student-teacher setting in which an ove
r-parameterized student network learns from the output of a fixed teacher networ
k of the same depth, with Stochastic Gradient Descent (SGD). Our contributions a
re two-fold. First, we prove that when the gradient is zero (or bounded above by
 a small constant) at every data point in training, a situation called \emph{in
terpolation setting}, there exists many-to-one \emph{alignment} between student
and teacher nodes in the lowest layer under mild conditions. This suggests that
generalization in unseen dataset is achievable, even the same condition often le
ads to zero training error. Second, analysis of noisy recovery and training dyna
mics in 2-layer network shows that strong teacher nodes (with large fan-out weig
hts) are learned first and subtle teacher nodes are left unlearned until late st
age of training. As a result, it could take a long time to converge into these s
mall-gradient critical points. Our analysis shows that over-parameterization pla
ys two roles: (1) it is a necessary condition for alignment to happen at the cri
tical points, and (2) in training dynamics, it helps student nodes cover more te
acher nodes with fewer iterations. Both improve generalization. Experiments just
ify our finding.
****

Asymptotics of Wide Networks from Feynman Diagrams
Ethan Dyer,Guy Gur-Ari
Understanding the asymptotic behavior of wide networks is of considerable intere
st. In this work, we present a general method for analyzing this large width beh
avior. The method is an adaptation of Feynman diagrams, a standard tool for comp
uting multivariate Gaussian integrals. We apply our method to study training dyn
amics, improving existing bounds and deriving new results on wide network evolut
ion during stochastic gradient descent. Going beyond the strict large width limi
t, we present closed-form expressions for higher-order terms governing wide netw
ork training, and test these predictions empirically.

****

Symplectic Recurrent Neural Networks
Zhengdao Chen,Jianyu Zhang,Martin Arjovsky,Léon Bottou
We propose Symplectic Recurrent Neural Networks (SRNNs) as learning algorithms t

hat capture the dynamics of physical systems from observed trajectories. SRNNs m
odel the Hamiltonian function of the system by a neural networks, and leverage s
ymplectic integration, multiple-step training and initial state optimization to
address the challenging numerical issues associated with Hamiltonian systems. We
 show SRNNs succeed reliably on complex and noisy Hamiltonian systems. Finally,
we show how to augment the SRNN integration scheme in order to handle stiff dyna
mical systems such as bouncing billiards.
****
Generalized Bayesian Posterior Expectation Distillation for Deep Neural Networks
Meet P. Vadera,Benjamin M. Marlin
In this paper, we present a general framework for distilling expectations with r
espect to the Bayesian posterior distribution of a deep neural network, signific
antly extending prior work on a method known as ``Bayesian Dark Knowledge." Our
 generalized framework applies to the case of classification models and takes as
 input the architecture of a ``teacher" network, a general posterior expectation
 of interest, and the architecture of a ``student" network. The distillation met
hod performs an online compression of the selected posterior expectation using i
teratively generated Monte Carlo samples from the parameter posterior of the tea
cher model. We further consider the problem of optimizing the student model arch
itecture with respect to an accuracy-speed-storage trade-off. We present experim
ental results investigating multiple data sets, distillation targets, teacher m
odel architectures, and approaches to searching for student model architectures.
 We establish the key result that distilling into a student model with an archit
ecture that matches the teacher, as is done in Bayesian Dark Knowledge, can lead
 to sub-optimal performance. Lastly, we show that student architecture search me
thods can identify student models with significantly improved performance.
****
Learning Cross-Context Entity Representations from Text
Jeffrey Ling,Nicholas FitzGerald,Zifei Shan,Livio Baldini Soares,Thibault Févry,
David Weiss,Tom Kwiatkowski
Language modeling tasks, in which words, or word-pieces, are predicted on the ba
sis of a local context, have been very effective for learning word embeddings an
d context dependent representations of phrases. Motivated by the observation tha
t efforts to code world knowledge into machine readable knowledge bases or human
 readable encyclopedias tend to be entity-centric, we investigate the use of a f
ill-in-the-blank task to learn context independent representations of entities f
rom the text contexts in which those entities were mentioned. We show that large
 scale training of neural models allows us to learn high quality entity represen
tations, and we demonstrate successful results on four domains: (1) existing ent
ity-level typing benchmarks, including a 64% error reduction over previous work
on TypeNet (Murty et al., 2018); (2) a novel few-shot category reconstruction ta
sk; (3) existing entity linking benchmarks, where we achieve a score of 87.3% on
 TAC-KBP 2010 without using any alias table, external knowledge base or in domai
n training data and (4) answering trivia questions, which uniquely identify enti
ties. Our global entity representations encode fine-grained type categories, suc
h as "Scottish footballers", and can answer trivia questions such as "Who was th
e last inmate of Spandau jail in Berlin?".
****
SPECTRA: Sparse Entity-centric Transitions
Rim Assouel,Yoshua Bengio
Learning an agent that interacts with objects is ubiquituous in many RL tasks. I
n most of them the agent's actions have sparse effects : only a small subset of
 objects in the visual scene will be affected by the action taken. We introduce
SPECTRA, a model for learning slot-structured transitions from raw visual observ
ations that embodies this sparsity assumption. Our model is composed of a percep
tion module that decomposes the visual scene into a set of latent objects repre
sentations (i.e. slot-structured) and a transition module that predicts the next
 latent set slot-wise and in a sparse way. We show that learning a perception mo
dule jointly with a sparse slot-structured transition model not only biases the
 model towards more entity-centric perceptual groupings but also enables intrin

sic exploration strategy that aims at maximizing the number of objects changed i
n the agent’s trajectory.
****

DeepSimplex: Reinforcement Learning of Pivot Rules Improves the Efficiency of Si
mplex Algorithm in Solving Linear Programming Problems

Varun Suriyanarayana,Onur Tavaslioglu,Ankit B. Patel,Andrew J. Schaefer

Linear Programs (LPs) are a fundamental class of optimization problems with a wi
de variety of applications. Fast algorithms for solving LPs are the workhorse of
 many combinatorial optimization algorithms, especially those involving integer
programming. One popular method to solve LPs is the simplex method which, at eac
h iteration, traverses the surface of the polyhedron of feasible solutions. At e
ach vertex of the polyhedron, one of several heuristics chooses the next neighbo
ring vertex, and these vary in accuracy and computational cost. We use deep valu
e-based reinforcement learning to learn a pivoting strategy that at each iterati
on chooses between two of the most popular pivot rules -- Dantzig and steepest e
dge.
Because the latter is typically more accurate and computationally costly than th
e former, we assign a higher wall time-based cost to steepest edge iterations th
an Dantzig iterations. We optimize this weighted cost on a neural net architectu
re designed for the simplex algorithm. We obtain between 20% to 50% reduction in
 the gap between weighted iterations of the individual pivoting rules, and the b
est possible omniscient policies for LP relaxations of randomly generated instan
ces of five-city Traveling Salesman Problem.
****

Learning Temporal Abstraction with Information-theoretic Constraints for Hierarc
hical Reinforcement Learning

Wenshan Wang,Yaoyu Hu,Sebastian Scherer

Applying reinforcement learning (RL) to real-world problems will require reasoni
ng about action-reward correlation over long time horizons. Hierarchical reinfor
cement learning (HRL) methods handle this by dividing the task into hierarchies,
 often with hand-tuned network structure or pre-defined subgoals. We propose a n
ovel HRL framework TAIC, which learns the temporal abstraction from past experie
nce or expert demonstrations without task-specific knowledge. We formulate the t
emporal abstraction problem as learning latent representations of action sequenc
es and present a novel approach of regularizing the latent space by adding infor
mation-theoretic constraints. Specifically, we maximize the mutual information b
etween the latent variables and the state changes.
A visualization of the latent space demonstrates that our algorithm learns an ef
fective abstraction of the long action sequences. The learned abstraction allows
 us to learn new tasks on higher level more efficiently. We convey a significant
 speedup in convergence over benchmark learning problems. These results demonstr
ate that learning temporal abstractions is an effective technique in increasing
the convergence rate and sample efficiency of RL algorithms.
****

Selective Brain Damage: Measuring the Disparate Impact of Model Pruning

Sara Hooker,Yann Dauphin,Aaron Courville,Andrea Frome

Neural network pruning techniques have demonstrated it is possible to remove the
 majority of weights in a network with surprisingly little degradation to top-1
test set accuracy. However, this measure of performance conceals significant dif
ferences in how different classes and images are impacted by pruning. We find th
at certain individual data points, which we term pruning identified exemplars (P
IEs), and classes are systematically more impacted by the introduction of sparsi
ty. Removing PIE images from the test-set greatly improves top-1 accuracy for bo
th sparse and non-sparse models. These hard-to-generalize-to images tend to be o
f lower image quality, mislabelled, entail abstract representations, require fin
e-grained classification or depict atypical class examples.
****

Asynchronous Stochastic Subgradient Methods for General Nonsmooth Nonconvex Opti
mization

Vyacheslav Kungurtsev,Malcolm Egan,Bapi Chatterjee,Dan Alistarh

Asynchronous distributed methods are a popular way to reduce the communication a
nd synchronization costs of large-scale optimization. Yet, for all their success
, little is known about their convergence guarantees in the challenging case of
general non-smooth, non-convex objectives, beyond cases where closed-form proxim
al operator solutions are available.
This is all the more surprising since these objectives are the ones appearing in
 the training of deep neural networks.

In this paper, we introduce the first convergence analysis covering asynchronous
 methods in the case of general non-smooth, non-convex objectives. Our analysis
applies to stochastic sub-gradient descent methods both with and without block v
ariable partitioning, and both with and without momentum. It is phrased in the c
ontext of a general probabilistic model of asynchronous scheduling accurately ad
apted to modern hardware properties. We validate our analysis experimentally in
the context of training deep neural network architectures. We show their overall
 successful asymptotic convergence as well as exploring how momentum, synchroniz
ation, and partitioning all affect performance.
****
Improved Structural Discovery and Representation Learning of Multi-Agent Data
Jennifer Hobbs,Matthew Holbrook,Nathan Frank,Long Sha,Patrick Lucey
Central to all machine learning algorithms is data representation. For multi-age
nt systems, selecting a representation which adequately captures the interaction
s among agents is challenging due to the latent group structure which tends to v
ary depending on various contexts. However, in multi-agent systems with strong g
roup structure, we can simultaneously learn this structure and map a set of agen
ts to a consistently ordered representation for further learning. In this paper,
 we present a dynamic alignment method which provides a robust ordering of struc
tured multi-agent data which allows for representation learning to occur in a fr
action of the time of previous methods. We demonstrate the value of this approa
ch using a large amount of soccer tracking data from a professional league.
****
Quantized Reinforcement Learning (QuaRL)
Srivatsan Krishnan,Sharad Chitlangia,Maximilian Lam,Zishen Wan,Aleksandra Faust,
Vijay Janapa Reddi
Recent work has shown that quantization can help reduce the memory, compute, and
 energy demands of deep neural networks without significantly harming their qual
ity. However, whether these prior techniques, applied traditionally to image-bas
ed models, work with the same efficacy to the sequential decision making process
 in reinforcement learning remains an unanswered question. To address this void,
 we conduct the first comprehensive empirical study that quantifies the effects
of quantization on various deep reinforcement learning policies with the intent
to reduce their computational resource demands. We apply techniques such as post
-training quantization and quantization aware training to a spectrum of reinforc
ement learning tasks (such as Pong, Breakout, BeamRider and more) and training a
lgorithms (such as PPO, A2C, DDPG, and DQN). Across this spectrum of tasks and l
earning algorithms, we show that policies can be quantized to 6-8 bits of precis
ion without loss of accuracy. Additionally, we show that certain tasks and reinf
orcement learning algorithms yield policies that are more difficult to quantize
due to their effect of widening the models' distribution of weights and that qua
ntization aware training consistently improves results over post-training quanti
zation and oftentimes even over the full precision baseline. Finally, we demonst
rate the real-world applications of quantization for reinforcement learning. We
use half-precision training to train a Pong model 50 % faster, and we deploy a q
uantized reinforcement learning based navigation policy to an embedded system, a
chieving an 18x speedup and a 4x reduction in memory usage over an unquantized p
olicy.
****
R-TRANSFORMER: RECURRENT NEURAL NETWORK ENHANCED TRANSFORMER
Zhiwei Wang,Yao Ma,Zitao Liu,Jiliang Tang
Recurrent Neural Networks have long been the dominating choice for sequence mode

ling. However, it severely suffers from two issues: impotent in capturing very l
ong-term dependencies and unable to parallelize the sequential computation proce
dure. Therefore, many non-recurrent sequence models that are built on convolutio
n and attention operations have been proposed recently. Notably, models with mul
ti-head attention such as Transformer have demonstrated extreme effectiveness in
 capturing long-term dependencies in a variety of sequence modeling tasks. Despi
te their success, however, these models lack necessary components to model local
 structures in sequences and heavily rely on position embeddings that have limit
ed effects and require a considerable amount of design efforts. In this paper, w
e propose the R-Transformer which enjoys the advantages of both RNNs and the mul
ti-head attention mechanism while avoids their respective drawbacks. The propose
d model can effectively capture both local structures and global long-term depen
dencies in sequences without any use of position embeddings. We evaluate R-Trans
former through extensive experiments with data from a wide range of domains and
the empirical results show that R-Transformer outperforms the state-of-the-art m
ethods by a large margin in most of the tasks.
****

NADS: Neural Architecture Distribution Search for Uncertainty Awareness
Randy Ardywibowo,Shahin Boluki,Xinyu Gong,Zhangyang Wang,Xiaoning Qian
Machine learning systems often encounter Out-of-Distribution (OoD) errors when d
ealing with testing data coming from a different distribution from the one used
for training. With their growing use in critical applications, it becomes import
ant to develop systems that are able to accurately quantify its predictive uncer
tainty and screen out these anomalous inputs. However, unlike standard learning
tasks, there is currently no well established guiding principle for designing ar
chitectures that can accurately quantify uncertainty. Moreover, commonly used Oo
D detection approaches are prone to errors and even sometimes assign higher like
lihoods to OoD samples. To address these problems, we first seek to identify gui
ding principles for designing uncertainty-aware architectures, by proposing Neur
al Architecture Distribution Search (NADS). Unlike standard neural architecture
search methods which seek for a single best performing architecture, NADS search
es for a distribution of architectures that perform well on a given task, allowi
ng us to identify building blocks common among all uncertainty aware architectur
es. With this formulation, we are able to optimize a stochastic outlier detectio
n objective and construct an ensemble of models to perform OoD detection. We per
form multiple OoD detection experiments and observe that our NADS performs favor
ably compared to state-of-the-art OoD detection methods.
****

Rigging the Lottery: Making All Tickets Winners
Utku Evci,Erich Elsen,Pablo Castro,Trevor Gale
Sparse neural networks have been shown to yield computationally efficient networ
ks with improved inference times. There is a large body of work on training den
se networks to yield sparse networks for inference (Molchanov et al., 2017;Zhu &
 Gupta, 2018; Louizos et al., 2017; Li et al., 2016; Guo et al., 2016). This li
mits the size of the largest trainable sparse model to that of the largest train
able dense model. In this paper we introduce a method to train sparse neural net
works with a fixed parameter count and a fixed computational cost throughout tra
ining, without sacrificing accuracy relative to existing dense-to-sparse trainin
g methods. Our method updates the topology of the network during training by usi
ng parameter magnitudes and infrequent gradient calculations. We show that this
approach requires less floating-point operations (FLOPs) to achieve a given leve
l of accuracy compared to prior techniques. We demonstrate state-of-the-art spar
se training results with ResNet-50, MobileNet v1 and MobileNet v2 on the ImageNe
t-2012 dataset. Finally, we provide some insights into why allowing the
 topology to change during the optimization can overcome local minima encounter
ed when the topology remains static.
****

CAPACITY-LIMITED REINFORCEMENT LEARNING: APPLICATIONS IN DEEP ACTOR-CRITIC METHO
DS FOR CONTINUOUS CONTROL
Tyler James Malloy,Matthew Riemer,Miao Liu,Tim Klinger,Gerald Tesauro,Chris R. S

ims
Biological and artificial agents must learn to act optimally in spite of a limited capacity for processing, storing, and attending to information. We formalize this type of bounded rationality in terms of an information-theoretic constraint on the complexity of policies that agents seek to learn. We present the Capacity-Limited Reinforcement Learning (CLRL) objective which defines an optimal policy subject to an information capacity constraint. This objective is optimized by drawing from methods used in rate distortion theory and information theory, and applied to the reinforcement learning setting. Using this objective we implement a novel Capacity-Limited Actor-Critic (CLAC) algorithm and situate it within a broader family of RL algorithms such as the Soft Actor Critic (SAC) and discuss their similarities and differences. Our experiments show that compared to alternative approaches, CLAC offers improvements in generalization between training and modified test environments. This is achieved in the CLAC model while displaying high sample efficiency and minimal requirements for hyper-parameter tuning.
****

Discovering the compositional structure of vector representations with Role Learning Networks
Paul Soulos,Tom McCoy,Tal Linzen,Paul Smolensky
Neural networks (NNs) are able to perform tasks that rely on compositional structure even though they lack obvious mechanisms for representing this structure. To analyze the internal representations that enable such success, we propose ROLE, a technique that detects whether these representations implicitly encode symbolic structure. ROLE learns to approximate the representations of a target encoder E by learning a symbolic constituent structure and an embedding of that structure into E’s representational vector space. The constituents of the approximating symbol structure are defined by structural positions — roles — that can be filled by symbols. We show that when E is constructed to explicitly embed a particular type of structure (e.g., string or tree), ROLE successfully extracts the ground-truth roles defining that structure. We then analyze a seq2seq network trained to perform a more complex compositional task (SCAN), where there is no ground truth role scheme available. For this model, ROLE successfully discovers an interpretable symbolic structure that the model implicitly uses to perform the SCAN task, providing a comprehensive account of the link between the representations and the behavior of a notoriously hard-to-interpret type of model. We verify the causal importance of the discovered symbolic structure by showing that, when we systematically manipulate hidden embeddings based on this symbolic structure, the model’s output is also changed in the way predicted by our analysis. Finally, we use ROLE to explore whether popular sentence embedding models are capturing compositional structure and find evidence that they are not; we conclude by discussing how insights from ROLE can be used to impart new inductive biases that will improve the compositional abilities of such models.
****

Higher-Order Function Networks for Learning Composable 3D Object Representations
Eric Mitchell,Selim Engin,Volkan Isler,Daniel D Lee
We present a new approach to 3D object representation where a neural network encodes the geometry of an object directly into the weights and biases of a second 'mapping' network. This mapping network can be used to reconstruct an object by applying its encoded transformation to points randomly sampled from a simple geometric space, such as the unit sphere. We study the effectiveness of our method through various experiments on subsets of the ShapeNet dataset. We find that the proposed approach can reconstruct encoded objects with accuracy equal to or exceeding state-of-the-art methods with orders of magnitude fewer parameters. Our smallest mapping network has only about 7000 parameters and shows reconstruction quality on par with state-of-the-art object decoder architectures with millions of parameters. Further experiments on feature mixing through the composition of learned functions show that the encoding captures a meaningful subspace of objects.
****

Adapting to Label Shift with Bias-Corrected Calibration

Avanti Shrikumar,Amr M. Alexandari,Anshul Kundaje
Label shift refers to the phenomenon where the marginal probability p(y) of obse rving a particular class changes between the training and test distributions, wh ile the conditional probability p(x|y) stays fixed. This is relevant in settings such as medical diagnosis, where a classifier trained to predict disease based on observed symptoms may need to be adapted to a different distribution where th e baseline frequency of the disease is higher. Given estimates of p(y|x) from a predictive model, one can apply domain adaptation procedures including Expectati on Maximization (EM) and Black-Box Shift Estimation (BBSE) to efficiently correc t for the difference in class proportions between the training and test distribu tions. Unfortunately, modern neural networks typically fail to produce well-cali brated estimates of p(y|x), reducing the effectiveness of these approaches. In r ecent years, Temperature Scaling has emerged as an efficient approach to combat miscalibration. However, the effectiveness of Temperature Scaling in the context of adaptation to label shift has not been explored. In this work, we study the impact of various calibration approaches on shift estimates produced by EM or BB SE. In experiments with image classification and diabetic retinopathy detection, we find that calibration consistently tends to improve shift estimation. In par ticular, calibration approaches that include class-specific bias parameters are significantly better than approaches that lack class-specific bias parameters, s uggesting that reducing systematic bias in the calibrated probabilities is espec ially important for domain adaptation.
****

Neural Module Networks for Reasoning over Text
Nitish Gupta,Kevin Lin,Dan Roth,Sameer Singh,Matt Gardner
Answering compositional questions that require multiple steps of reasoning again st text is challenging, especially when they involve discrete, symbolic operatio ns. Neural module networks (NMNs) learn to parse such questions as executable pr ograms composed of learnable modules, performing well on synthetic visual QA dom ains. However, we find that it is challenging to learn these models for non-synt hetic questions on open-domain text, where a model needs to deal with the divers ity of natural language and perform a broader range of reasoning. We extend NMNs by: (a) introducing modules that reason over a paragraph of text, performing sy mbolic reasoning (such as arithmetic, sorting, counting) over numbers and dates in a probabilistic and differentiable manner; and (b) proposing an unsupervised auxiliary loss to help extract arguments associated with the events in text. Add itionally, we show that a limited amount of heuristically-obtained question prog ram and intermediate module output supervision provides sufficient inductive bia s for accurate learning. Our proposed model significantly outperforms state-of-t he-art models on a subset of the DROP dataset that poses a variety of reasoning challenges that are covered by our modules.
****

MANIFOLD FORESTS: CLOSING THE GAP ON NEURAL NETWORKS
Ronan Perry,Tyler M. Tomita,Jesse Patsolic,Benjamin Falk,Joshua Vogelstein
Decision forests (DF), in particular random forests and gradient boosting trees, have demonstrated state-of-the-art accuracy compared to other methods in many supervised learning scenarios. In particular, DFs dominate other methods in tabu lar data, that is, when the feature space is unstructured, so that the signal is invariant to permuting feature indices. However, in structured data lying on a manifold---such as images, text, and speech---neural nets (NN) tend to outperfo rm DFs. We conjecture that at least part of the reason for this is that the inpu t to NN is not simply the feature magnitudes, but also their indices (for exampl e, the convolution operation uses ``feature locality). In contrast, naive DF imp lementations fail to explicitly consider feature indices. A recently proposed DF approach demonstrates that DFs, for each node, implicitly sample a random matri x from some specific distribution. Here, we build on that to show that one can choose distributions in a manifold aware fashion. For example, for image classif ication, rather than randomly selecting pixels, one can randomly select contiguo us patches. We demonstrate the empirical performance of data living on three di fferent manifolds: images, time-series, and a torus. In all three cases, our Man

ifold Forest (Mf) algorithm empirically dominates other state-of-the-art approac
hes that ignore feature space structure, achieving a lower classification error
on all sample sizes. This dominance extends to the MNIST data set as well. Moreo
ver, both training and test time is significantly faster for manifold forests as
 compared to deep nets. This approach, therefore, has promise to enable DFs and
other machine learning methods to close the gap with deep nets on manifold-value
d data.
****

Learning to Plan in High Dimensions via Neural Exploration-Exploitation Trees
Binghong Chen,Bo Dai,Qinjie Lin,Guo Ye,Han Liu,Le Song
We propose a meta path planning algorithm named \emph{Neural Exploration-Exploit
ation Trees~(NEXT)} for learning from prior experience for solving new path plan
ning problems in high dimensional continuous state and action spaces. Compared t
o more classical sampling-based methods like RRT, our approach achieves much bet
ter sample efficiency in high-dimensions and can benefit from prior experience
of planning in similar environments. More specifically, NEXT exploits a novel ne
ural architecture which can learn promising search directions from problem struc
tures. The learned prior is then integrated into a UCB-type algorithm to achieve
 an online balance between \emph{exploration} and \emph{exploitation} when solvi
ng a new problem. We conduct thorough experiments to show that NEXT accomplishes
 new planning problems with more compact search trees and significantly outperfo
rms state-of-the-art methods on several benchmarks.
****

Improved memory in recurrent neural networks with sequential non-normal dynamics
Emin Orhan,Xaq Pitkow
Training recurrent neural networks (RNNs) is a hard problem due to degeneracies
in the optimization landscape, a problem also known as vanishing/exploding gradi
ents. Short of designing new RNN architectures, previous methods for dealing wit
h this problem usually boil down to orthogonalization of the recurrent dynamics,
 either at initialization or during the entire training period. The basic motiva
tion behind these methods is that orthogonal transformations are isometries of t
he Euclidean space, hence they preserve (Euclidean) norms and effectively deal w
ith vanishing/exploding gradients. However, this ignores the crucial effects of
non-linearity and noise. In the presence of a non-linearity, orthogonal transfor
mations no longer preserve norms, suggesting that alternative transformations mi
ght be better suited to non-linear networks. Moreover, in the presence of noise,
 norm preservation itself ceases to be the ideal objective. A more sensible obje
ctive is maximizing the signal-to-noise ratio (SNR) of the propagated signal ins
tead. Previous work has shown that in the linear case, recurrent networks that m
aximize the SNR display strongly non-normal, sequential dynamics and orthogonal
networks are highly suboptimal by this measure. Motivated by this finding, here
we investigate the potential of non-normal RNNs, i.e. RNNs with a non-normal rec
urrent connectivity matrix, in sequential processing tasks. Our experimental res
ults show that non-normal RNNs outperform their orthogonal counterparts in a div
erse range of benchmarks. We also find evidence for increased non-normality and
hidden chain-like feedforward motifs in trained RNNs initialized with orthogonal
 recurrent connectivity matrices.
****

Model Imitation for Model-Based Reinforcement Learning
Yueh-Hua Wu,Ting-Han Fan,Peter J. Ramadge,Hao Su
Model-based reinforcement learning (MBRL) aims to learn a dynamic model to reduc
e the number of interactions with real-world environments. However, due to estim
ation error, rollouts in the learned model, especially those of long horizon, fa
il to match the ones in real-world environments. This mismatching has seriously
impacted the sample complexity of MBRL. The phenomenon can be attributed to the
fact that previous works employ supervised learning to learn the one-step transi
tion models, which has inherent difficulty ensuring the matching of distribution
s from multi-step rollouts. Based on the claim, we propose to learn the synthesi
zed model by matching the distributions of multi-step rollouts sampled from the
synthesized model and the real ones via WGAN. We theoretically show that matchin

g the two can minimize the difference of cumulative rewards between the real tra nsition and the learned one. Our experiments also show that the proposed model i mitation method outperforms the state-of-the-art in terms of sample complexity a nd average return.
****
Likelihood Contribution based Multi-scale Architecture for Generative Flows
Hari Prasanna Das,Pieter Abbeel,Costas J. Spanos
Deep generative modeling using flows has gained popularity owing to the tractabl e exact log-likelihood estimation with efficient training and synthesis process. However, flow models suffer from the challenge of having high dimensional laten t space, same in dimension as the input space. An effective solution to the abov e challenge as proposed by Dinh et al. (2016) is a multi-scale architecture, whi ch is based on iterative early factorization of a part of the total dimensions a t regular intervals. Prior works on generative flows involving a multi-scale arc hitecture perform the dimension factorization based on a static masking. We prop ose a novel multi-scale architecture that performs data dependent factorization to decide which dimensions should pass through more flow layers. To facilitate t he same, we introduce a heuristic based on the contribution of each dimension to the total log-likelihood which encodes the importance of the dimensions. Our pr oposed heuristic is readily obtained as part of the flow training process, enabl ing versatile implementation of our likelihood contribution based multi-scale ar chitecture for generic flow models. We present such an implementation for the or iginal flow introduced in Dinh et al. (2016), and demonstrate improvements in lo g-likelihood score and sampling quality on standard image benchmarks. We also co nduct ablation studies to compare proposed method with other options for dimensi on factorization.
****
A Base Model Selection Methodology for Efficient Fine-Tuning
Yosuke Ueno,Masaaki Kondo
While the accuracy of image classification achieves significant improvement with deep Convolutional Neural Networks (CNN), training a deep CNN is a time-consumi ng task because it requires a large amount of labeled data and takes a long time to converge even with high performance computing resources.
Fine-tuning, one of the transfer learning methods, is effective in decreasing ti me and the amount of data necessary for CNN training. It is known that fine-tuni ng can be performed efficiently if the source and the target tasks have high rel ativity.
However, the technique to evaluate the relativity or transferability of trained models quantitatively from their parameters has not been established. In this pa per, we propose and evaluate several metrics to estimate the transferability of pre-trained CNN models for a given target task by featuremaps of the last convol utional layer.
We found that some of the proposed metrics are good predictors of fine-tuned acc uracy, but their effectiveness depends on the structure of the network. Therefor e, we also propose to combine two metrics to get a generally applicable indicato r.
The experimental results reveal that one of the combined metrics is well correla ted with fine-tuned accuracy in a variety of network structure and our method ha s a good potential to reduce the burden of CNN training.
****
Rethinking Curriculum Learning With Incremental Labels And Adaptive Compensation
Madan Ravi Ganesh,Jason J. Corso
Like humans, deep networks learn better when samples are organized and introduce d in a meaningful order or curriculum. While conventional approaches to curricul um learning emphasize the difficulty of samples as the core incremental strategy , it forces networks to learn from small subsets of data while introducing pre-c omputation overheads. In this work, we propose Learning with Incremental Labels and Adaptive Compensation (LILAC), which introduces a novel approach to curricul um learning. LILAC emphasizes incrementally learning labels instead of increment ally learning difficult samples. It works in two distinct phases: first, in the

incremental label introduction phase, we unmask ground-truth labels in fixed inc
rements during training, to improve the starting point from which networks learn
. In the adaptive compensation phase, we compensate for failed predictions by ad
aptively altering the target vector to a smoother distribution. We evaluate LILA
C against the closest comparable methods in batch and curriculum learning and la
bel smoothing, across three standard image benchmarks, CIFAR-10, CIFAR-100, and
STL-10. We show that our method outperforms batch learning with higher mean reco
gnition accuracy as well as lower standard deviation in performance consistently
 across all benchmarks. We further extend LILAC to state-of-the-art performance
across CIFAR-10 using simple data augmentation while exhibiting label order inva
riance among other important properties.
****
Graph Neural Networks for Reasoning 2-Quantified Boolean Formulas
Fei Wang,Zhanfu Yang,Ziliang Chen,Guannan Wei,Tiark Rompf
It is valuable yet remains challenging to apply neural networks in logical reaso
ning tasks. Despite some successes witnessed in learning SAT (Boolean Satisfiabi
lity) solvers for propositional logic via Graph Neural Networks (GNN), there ha
ven't been any successes in learning solvers for more complex predicate logic. I
n this paper, we target the QBF (Quantified Boolean Formula) satisfiability prob
lem, the complexity of which is in-between propositional logic and predicate log
ic, and investigate the feasibility of learning GNN-based solvers and GNN-based
heuristics for the cases with a universal-existential quantifier alternation (so
-called 2QBF problems).

We conjecture, with empirical support, that GNNs have certain limitations in lea
rning 2QBF solvers, primarily due to the inability to reason about a set of assi
gnments. Then we show the potential of GNN-based heuristics in CEGAR-based solve
rs and explore the interesting challenges to generalize them to larger problem i
nstances. In summary, this paper provides a comprehensive surveying view of appl
ying GNN-based embeddings to 2QBF problems and aims to offer insights in applyin
g machine learning tools to more complicated symbolic reasoning problems.

****
Learn to Explain Efficiently via Neural Logic Inductive Learning
Yuan Yang,Le Song
The capability of making interpretable and self-explanatory decisions is essenti
al for developing responsible machine learning systems. In this work, we study t
he learning to explain the problem in the scope of inductive logic programming (
ILP). We propose Neural Logic Inductive Learning (NLIL), an efficient differenti
able ILP framework that learns first-order logic rules that can explain the patt
erns in the data. In experiments, compared with the state-of-the-art models, we
find NLIL is able to search for rules that are x10 times longer while remaining
x3 times faster. We also show that NLIL can scale to large image datasets, i.e.
Visual Genome, with 1M entities.
****
NormLime: A New Feature Importance Metric for Explaining Deep Neural Networks
Isaac Ahern,Adam Noack,Luis Guzman-Nateras,Dejing Dou,Boyang Li,Jun Huan
The problem of explaining deep learning models, and model predictions generally,
 has attracted intensive interest recently. Many successful approaches forgo glo
bal approximations in order to provide more faithful local interpretations of th
e model’s behavior. LIME develops multiple interpretable models, each approxima
ting a large neural network on a small region of the data manifold, and SP-LIME
aggregates the local models to form a global interpretation. Extending this lin
e of research, we propose a simple yet effective method, NormLIME, for aggregati
ng local models into global and class-specific interpretations. A human user st
udy strongly favored the class-specific interpretations created by NormLIME to o
ther feature importance metrics. Numerical experiments employing Keep And Retrai
n (KAR) based feature ablation across various baselines (Random, Gradient-based,
 LIME, SHAP) confirms NormLIME’s effectiveness for recognizing important feature
s.

****
Pre-trained Contextual Embedding of Source Code
Aditya Kanade,Petros Maniatis,Gogul Balakrishnan,Kensen Shi
The source code of a program not only serves as a formal description of an executable task, but it also serves to communicate developer intent in a human-readable form. To facilitate this, developers use meaningful identifier names and natural-language documentation. This makes it possible to successfully apply sequence-modeling approaches, shown to be effective in natural-language processing, to source code. A major advancement in natural-language understanding has been the use of pre-trained token embeddings; BERT and other works have further shown that pre-trained contextual embeddings can be extremely powerful and can be finetuned effectively for a variety of downstream supervised tasks. Inspired by these developments, we present the first attempt to replicate this success on source code. We curate a massive corpus of Python programs from GitHub to pre-train a BERT model, which we call Code Understanding BERT (CuBERT). We also pre-train Word2Vec embeddings on the same dataset. We create a benchmark of five classification tasks and compare finetuned CuBERT against sequence models trained with and without the Word2Vec embeddings. Our results show that CuBERT outperforms the baseline methods by a margin of 2.9-22%. We also show its superiority when finetuned with smaller datasets, and over fewer epochs.
****
Certified Robustness to Adversarial Label-Flipping Attacks via Randomized Smoothing
Elan Rosenfeld,Ezra Winston,Pradeep Ravikumar,J. Zico Kolter
This paper considers label-flipping attacks, a type of data poisoning attack where an adversary relabels a small number of examples in a training set in order to degrade the performance of the resulting classifier. In this work, we propose a strategy to build classifiers that are certifiably robust against a strong variant of label-flipping, where the adversary can target each test example independently. In other words, for each test point, our classifier makes a prediction and includes a certification that its prediction would be the same had some number of training labels been changed adversarially. Our approach leverages randomized smoothing, a technique that has previously been used to guarantee test-time robustness to adversarial manipulation of the input to a classifier. Further, we obtain these certified bounds with no additional runtime cost over standard classification. On the Dogfish binary classification task from ImageNet, in the face of an adversary who is allowed to flip 10 labels to individually target each test point, the baseline undefended classifier achieves no more than 29.3% accuracy; we obtain a classifier that maintains 64.2% certified accuracy against the same adversary.
****
Benefit of Interpolation in Nearest Neighbor Algorithms
Yue Xing,Qifan Song,Guang Cheng
The over-parameterized models attract much attention in the era of data science and deep learning. It is empirically observed that although these models, e.g. deep neural networks, over-fit the training data, they can still achieve small testing error, and sometimes even outperform traditional algorithms which are designed to avoid over-fitting. The major goal of this work is to sharply quantify the benefit of data interpolation in the context of nearest neighbors (NN) algorithm. Specifically, we consider a class of interpolated weighting schemes and then carefully characterize their asymptotic performances. Our analysis reveals a U-shaped performance curve with respect to the level of data interpolation, and proves that a mild degree of data interpolation strictly improves the prediction accuracy and statistical stability over those of the (un-interpolated) optimal kNN algorithm. This theoretically justifies (predicts) the existence of the second U-shaped curve in the recently discovered double descent phenomenon. Note that our goal in this study is not to promote the use of interpolated-NN method, but to obtain theoretical insights on data interpolation inspired by the aforementioned phenomenon.
****

{COMPANYNAME}11K: An Unsupervised Representation Learning Dataset for Arrhythmia Subtype Discovery

Shawn Tan,Guillaume Androz,Ahmad Chamseddine,Pierre Fecteau,Aaron Courville,Yoshua Bengio,Joseph Paul Cohen

We release the largest public ECG dataset of continuous raw signals for representation learning containing over 11k patients and 2 billion labelled beats. Our goal is to enable semi-supervised ECG models to be made as well as to discover unknown subtypes of arrhythmia and anomalous ECG signal events. To this end, we propose an unsupervised representation learning task, evaluated in a semi-supervised fashion. We provide a set of baselines for different feature extractors that can be built upon. Additionally, we perform qualitative evaluations on results from PCA embeddings, where we identify some clustering of known subtypes indicating the potential for representation learning in arrhythmia sub-type discovery.

****

Neural Clustering Processes

Ari Pakman,Yueqi Wang,Catalin Mitelut,JinHyung Lee,Liam Paninski

Mixture models, a basic building block in countless statistical models, involve latent random variables over discrete spaces, and existing posterior inference methods can be inaccurate and/or very slow. In this work we introduce a novel deep learning architecture for efficient amortized Bayesian inference over mixture models. While previous approaches to amortized clustering assumed a fixed or maximum number of mixture components and only amortized over the continuous parameters of each mixture component, our method amortizes over the local discrete labels of all the data points, and performs inference over an unbounded number of mixture components. The latter property makes our method natural for the challenging case of nonparametric Bayesian models, where the number of mixture components grows with the dataset. Our approach exploits the exchangeability of the generative models and is based on mapping distributed, permutation-invariant representations of discrete arrangements into varying-size multinomial conditional probabilities. The resulting algorithm parallelizes easily, yields iid samples from the approximate posteriors along with a normalized probability estimate of each sample (a quantity generally unavailable using Markov Chain Monte Carlo) and can easily be applied to both conjugate and non-conjugate models, as training only requires samples from the generative model. We also present an extension of the method to models of random communities (such as infinite relational or stochastic block models). As a scientific application, we present a novel approach to neural spike sorting for high-density multielectrode arrays.

****

Improving Neural Language Generation with Spectrum Control

Lingxiao Wang,Jing Huang,Kevin Huang,Ziniu Hu,Guangtao Wang,Quanquan Gu

Recent Transformer-based models such as Transformer-XL and BERT have achieved huge success on various natural language processing tasks. However, contextualized embeddings at the output layer of these powerful models tend to degenerate and occupy an anisotropic cone in the vector space, which is called the representation degeneration problem. In this paper, we propose a novel spectrum control approach to address this degeneration problem. The core idea of our method is to directly guide the spectra training of the output embedding matrix with a slow-decaying singular value prior distribution through a reparameterization framework. We show that our proposed method encourages isotropy of the learned word representations while maintains the modeling power of these contextual neural models. We further provide a theoretical analysis and insight on the benefit of modeling singular value distribution. We demonstrate that our spectrum control method outperforms the state-of-the-art Transformer-XL modeling for language model, and various Transformer-based models for machine translation, on common benchmark datasets for these tasks.

****

Span Recovery for Deep Neural Networks with Applications to Input Obfuscation

Rajesh Jayaram,David P. Woodruff,Qiuyi Zhang

The tremendous success of deep neural networks has motivated the need to better

understand the fundamental properties of these networks, but many of the theoret ical results proposed have only been for shallow networks. In this paper, we stu dy an important primitive for understanding the meaningful input space of a deep network: span recovery. For $k<n$, let $\mathbf{A} \in \mathbb{R}^{k \times n}$ be the innermost weight matrix of an arbitrary feed forward neural network $M: \mathbb{R}^n \to \mathbb{R}$, so $M(x)$ can be written as $M(x) = \sigma(\mathb f{A} x)$, for some network $\sigma: \mathbb{R}^k \to \mathbb{R}$. The goal is t hen to recover the row span of $\mathbf{A}$ given only oracle access to the valu e of $M(x)$. We show that if M is a multi-layered network with ReLU activation functions, then partial recovery is possible: namely, we can provably recover $ k/2$ linearly independent vectors in the row span of $\mathbf{A}$ using poly$(n) $ non-adaptive queries to $M(x)$. Furthermore, if M has differentiable activa tion functions, we demonstrate that \textit{full} span recovery is possible even when the output is first passed through a sign or $0/1$ thresholding function; in this case our algorithm is adaptive. Empirically, we confirm that full span r ecovery is not always possible, but only for unrealistically thin layers. For re asonably wide networks, we obtain full span recovery on both random networks and networks trained on MNIST data. Furthermore, we demonstrate the utility of span recovery as an attack by inducing neural networks to misclassify data obfuscate d by controlled random noise as sensical inputs.

****

Unknown-Aware Deep Neural Network

Lei Cao,Yizhou Yan,Samuel Madden,Elke Rundensteiner

An important property of image classification systems in the real world is that they both accurately classify objects from target classes (``knowns'') and safel y reject unknown objects (``unknowns'') that belong to classes not present in th e training data. Unfortunately, although the strong generalization ability of ex isting CNNs ensures their accuracy when classifying known objects, it also cause s them to often assign an unknown to a target class with high confidence. As a r esult, simply using low-confidence detections as a way to detect unknowns does n ot work well. In this work, we propose an Unknown-aware Deep Neural Network (UDN for short) to solve this challenging problem. The key idea of UDN is to enhance existing CNNs to support a product operation that models the product relationsh ip among the features produced by convolutional layers. This way, missing a sing le key feature of a target class will greatly reduce the probability of assignin g an object to this class. UDN uses a learned ensemble of these product operatio ns, which allows it to balance the contradictory requirements of accurately clas sifying known objects and correctly rejecting unknowns. To further improve the p erformance of UDN at detecting unknowns, we propose an information-theoretic reg ularization strategy that incorporates the objective of rejecting unknowns into the learning process of UDN. We experiment on benchmark image datasets including MNIST, CIFAR-10, CIFAR-100, and SVHN, adding unknowns by injecting one dataset into another. Our results demonstrate that UDN significantly outperforms state-o f-the-art methods at rejecting unknowns by 25 percentage points improvement in a ccuracy, while still preserving the classification accuracy.

****

MODELLING BIOLOGICAL ASSAYS WITH ADAPTIVE DEEP KERNEL LEARNING

Prudencio Tossou,Basile Dura,Daniel Cohen,Mario Marchand,François Laviolette,Ale xandre Lacoste

Due to the significant costs of data generation, many prediction tasks within dr ug discovery are by nature few-shot regression (FSR) problems, including accurat e modelling of biological assays. Although a number of few-shot classification and reinforcement learning methods exist for similar applications, we find relat ively few FSR methods meeting the performance standards required for such tasks under real-world constraints. Inspired by deep kernel learning, we develop a nov el FSR algorithm that is better suited to these settings. Our algorithm consists of learning a deep network in combination with a kernel function and a differen tiable kernel algorithm. As the choice of the kernel is critical, our algorithm learns to find the appropriate one for each task during inference. It thus perfo

rms more effectively with complex task distributions, outperforming current stat e-of-the-art algorithms on both toy and novel, real-world benchmarks that we int roduce herein. By introducing novel benchmarks derived from biological assays, w e hope that the community will progress towards the development of FSR algorithm s suitable for use in noisy and uncertain environments such as drug discovery.
****

A Perturbation Analysis of Input Transformations for Adversarial Attacks
Adam Dziedzic,Sanjay Krishnan
The existence of adversarial examples, or intentional mis-predictions constructe d from small changes to correctly predicted examples, is one of the most signifi cant challenges in neural network research today. Ironically, many new defenses are based on a simple observation - the adversarial inputs themselves are not ro bust and small perturbations to the attacking input often recover the desired pr ediction. While the intuition is somewhat clear, a detailed understanding of thi s phenomenon is missing from the research literature. This paper presents a comp rehensive experimental analysis of when and why perturbation defenses work and p otential mechanisms that could explain their effectiveness (or ineffectiveness) in different settings.
****

ADA+: A GENERIC FRAMEWORK WITH MORE ADAPTIVE EXPLICIT ADJUSTMENT FOR LEARNING RA TE
Yue Zhao,Xiangsheng Huang,Ludan Kou
Although adaptive algorithms have achieved significant success in training deep neural networks with faster training speed, they tend to have poor generalizatio n performance compared to SGD with Momentum(SGDM). One of the state-of-the-art a lgorithms, PADAM, is proposed to close the generalization gap of adaptive method s while lacking an internal explanation. This work pro- poses a general framewor k, in which we use an explicit function $\Phi(\cdot)$ as an adjustment to the actual step size, and present a more adaptive specific form AdaPlus(Ada+). Based on this fr amework, we analyze various behaviors brought by different types of $\Phi(\cdot)$, such a s a constant function in SGDM, a linear function in Adam, a concave function in Padam and a concave function with offset term in AdaPlus. Empirically, we conduc t experiments on classic benchmarks both in CNN and RNN architectures and achiev e better performance(even than SGDM).

****

Oblique Decision Trees from Derivatives of ReLU Networks
Guang-He Lee,Tommi S. Jaakkola
We show how neural models can be used to realize piece-wise constant functions s uch as decision trees. The proposed architecture, which we call locally constant networks, builds on ReLU networks that are piece-wise linear and hence their as sociated gradients with respect to the inputs are locally constant. We formally establish the equivalence between the classes of locally constant networks and d ecision trees. Moreover, we highlight several advantageous properties of locally constant networks, including how they realize decision trees with parameter sha ring across branching / leaves. Indeed, only M neurons suffice to implicitly m odel an oblique decision tree with 2^M leaf nodes. The neural representation a lso enables us to adopt many tools developed for deep networks (e.g., DropConnec t (Wan et al., 2013)) while implicitly training decision trees. We demonstrate t hat our method outperforms alternative techniques for training oblique decision trees in the context of molecular property classification and regression tasks.
****

Smooth Kernels Improve Adversarial Robustness and Perceptually-Aligned Gradients
Haohan Wang,Xindi Wu,Songwei Ge,Zachary C. Lipton,Eric P. Xing
Recent research has shown that CNNs are often overly sensitive to high-frequency textural patterns. Inspired by the intuition that humans are more sensitive to the lower-frequency (larger-scale) patterns we design a regularization scheme th at penalizes large differences between adjacent components within each convoluti onal kernel. We apply our regularization onto several popular training methods, demonstrating that the models with the proposed smooth kernels enjoy improved ad

versarial robustness. Further, building on recent work establishing connections between adversarial robustness and interpretability, we show that our method appears to give more perceptually-aligned gradients.
****

Neural ODEs for Image Segmentation with Level Sets
Rafael Valle,Fitsum Reda,Mohammad Shoeybi,Patrick Legresley,Andrew Tao,Bryan Catanzaro
We propose a novel approach for image segmentation that combines Neural Ordinary Differential Equations (NODEs) and the Level Set method. Our approach parametrizes the evolution of an initial contour with a NODE that implicitly learns from data a speed function describing the evolution. In addition, for cases where an initial contour is not available and to alleviate the need for careful choice or design of contour embedding functions, we propose a NODE-based method that evolves an image embedding into a dense per-pixel semantic label space. We evaluate our methods on kidney segmentation (KiTS19) and on salient object detection (PASCAL-S, ECSSD and HKU-IS). In addition to improving initial contours provided by deep learning models while using a fraction of their number of parameters, our approach achieves F scores that are higher than several state-of-the-art deep learning algorithms
****

Precision Gating: Improving Neural Network Efficiency with Dynamic Dual-Precision Activations
Yichi Zhang,Ritchie Zhao,Weizhe Hua,Nayun Xu,G. Edward Suh,Zhiru Zhang
We propose precision gating (PG), an end-to-end trainable dynamic dual-precision quantization technique for deep neural networks. PG computes most features in a low precision and only a small proportion of important features in a higher precision to preserve accuracy. The proposed approach is applicable to a variety of DNN architectures and significantly reduces the computational cost of DNN execution with almost no accuracy loss. Our experiments indicate that PG achieves excellent results on CNNs, including statically compressed mobile-friendly networks such as ShuffleNet. Compared to the state-of-the-art prediction-based quantization schemes, PG achieves the same or higher accuracy with 2.4× less compute on ImageNet. PG furthermore applies to RNNs. Compared to 8-bit uniform quantization, PG obtains a 1.2% improvement in perplexity per word with 2.7× computational cost reduction on LSTM on the Penn Tree Bank dataset.
****

PAC Confidence Sets for Deep Neural Networks via Calibrated Prediction
Sangdon Park,Osbert Bastani,Nikolai Matni,Insup Lee
We propose an algorithm combining calibrated prediction and generalization bounds from learning theory to construct confidence sets for deep neural networks with PAC guarantees---i.e., the confidence set for a given input contains the true label with high probability. We demonstrate how our approach can be used to construct PAC confidence sets on ResNet for ImageNet, a visual object tracking model, and a dynamics model for the half-cheetah reinforcement learning problem.
****

Low Rank Training of Deep Neural Networks for Emerging Memory Technology
Albert Gural,Phillip Nadeau,Mehul Tikekar,Boris Murmann
The recent success of neural networks for solving difficult decision tasks has incentivized incorporating smart decision making "at the edge." However, this work has traditionally focused on neural network inference, rather than training, due to memory and compute limitations, especially in emerging non-volatile memory systems, where writes are energetically costly and reduce lifespan. Yet, the ability to train at the edge is becoming increasingly important as it enables applications such as real-time adaptability to device drift and environmental variation, user customization, and federated learning across devices. In this work, we address four key challenges for training on edge devices with non-volatile memory: low weight update density, weight quantization, low auxiliary memory, and online learning. We present a low-rank training scheme that addresses these four challenges while maintaining computational efficiency. We then demonstrate the technique on a representative convolutional neural network across several adaptati

on problems, where it out-performs standard SGD both in accuracy and in number o
f weight updates.
****
DD-PPO: Learning Near-Perfect PointGoal Navigators from 2.5 Billion Frames
Erik Wijmans,Abhishek Kadian,Ari Morcos,Stefan Lee,Irfan Essa,Devi Parikh,Manoli
s Savva,Dhruv Batra
We present Decentralized Distributed Proximal Policy Optimization (DD-PPO), a me
thod for distributed reinforcement learning in resource-intensive simulated envi
ronments. DD-PPO is distributed (uses multiple machines), decentralized (lacks a
 centralized server), and synchronous (no computation is ever "stale"), making i
t conceptually simple and easy to implement. In our experiments on training virt
ual robots to navigate in Habitat-Sim, DD-PPO exhibits near-linear scaling -- ac
hieving a speedup of 107x on 128 GPUs over a serial implementation. We leverage
this scaling to train an agent for 2.5 Billion steps of experience (the equivale
nt of 80 years of human experience) -- over 6 months of GPU-time training in und
er 3 days of wall-clock time with 64 GPUs.

This massive-scale training not only sets the state of art on Habitat Autonomous
 Navigation Challenge 2019, but essentially "solves" the task -- near-perfect au
tonomous navigation in an unseen environment without access to a map, directly f
rom an RGB-D camera and a GPS+Compass sensor. Fortuitously, error vs computatio
n exhibits a power-law-like distribution; thus, 90% of peak performance is obtai
ned relatively early (at 100 million steps) and relatively cheaply (under 1 day
with 8 GPUs). Finally, we show that the scene understanding and navigation polic
ies learned can be transferred to other navigation tasks -- the analog of "Image
Net pre-training + task-specific fine-tuning" for embodied AI. Our model outperf
orms ImageNet pre-trained CNNs on these transfer tasks and can serve as a univer
sal resource (all models and code are publicly available).
****
Learning to Learn by Zeroth-Order Oracle
Yangjun Ruan,Yuanhao Xiong,Sashank Reddi,Sanjiv Kumar,Cho-Jui Hsieh
In the learning to learn (L2L) framework, we cast the design of optimization alg
orithms as a machine learning problem and use deep neural networks to learn the
update rules. In this paper, we extend the L2L framework to zeroth-order (ZO) op
timization setting, where no explicit gradient information is available. Our lea
rned optimizer, modeled as recurrent neural network (RNN), first approximates gr
adient by ZO gradient estimator and then produces parameter update utilizing the
 knowledge of previous iterations. To reduce high variance effect due to ZO grad
ient estimator, we further introduce another RNN to learn the Gaussian sampling
rule and dynamically guide the query direction sampling. Our learned optimizer o
utperforms hand-designed algorithms in terms of convergence rate and final solut
ion on both synthetic and practical ZO optimization tasks (in particular, the bl
ack-box adversarial attack task, which is one of the most widely used tasks of Z
O optimization). We finally conduct extensive analytical experiments to demonstr
ate the effectiveness of our proposed optimizer.
****
Neural Embeddings for Nearest Neighbor Search Under Edit Distance
Xiyuan Zhang,Yang Yuan,Piotr Indyk
The edit distance between two sequences is an important metric with many applica
tions. The drawback, however, is the high computational cost of many basic probl
ems involving this notion, such as the nearest neighbor search. A natural approa
ch to overcoming this issue is to embed the sequences into a vector space such t
hat the geometric distance in the target space approximates the edit distance in
 the original space. However, the known edit distance embedding algorithms, such
 as Chakraborty et al.(2016), construct embeddings that are data-independent, i.
e., do not exploit any structure of embedded sets of strings. In this paper we p
ropose an alternative approach, which learns the embedding function according to
 the data distribution. Our experiments show that the new algorithm has much bet
ter empirical performance than prior data-independent methods.
****

ADAPTING PRETRAINED LANGUAGE MODELS FOR LONG DOCUMENT CLASSIFICATION
Matthew Lyle Olson,Lisa Zhang,Chun-Nam Yu
Pretrained language models (LMs) have shown excellent results in achieving human like performance on many language tasks. However, the most powerful LMs have one significant drawback: a fixed-sized input. With this constraint, these LMs are unable to utilize the full input of long documents. In this paper, we introduce a new framework to handle documents of arbitrary lengths. We investigate the addition of a recurrent mechanism to extend the input size and utilizing attention to identify the most discriminating segment of the input. We perform extensive validating experiments on patent and Arxiv datasets, both of which have long text. We demonstrate our method significantly outperforms state-of-the-art results reported in recent literature.
****
Robust Federated Learning Through Representation Matching and Adaptive Hyper-parameters
Hesham Mostafa
Federated learning is a distributed, privacy-aware learning scenario which trains a single model on data belonging to several clients. Each client trains a local model on its data and the local models are then aggregated by a central party. Current federated learning methods struggle in cases with heterogeneous client-side data distributions which can quickly lead to divergent local models and a collapse in performance. Careful hyper-parameter tuning is particularly important in these cases but traditional automated hyper-parameter tuning methods would require several training trials which is often impractical in a federated learning setting. We describe a two-pronged solution to the issues of robustness and hyper-parameter tuning in federated learning settings. We propose a novel representation matching scheme that reduces the divergence of local models by ensuring the feature representations in the global (aggregate) model can be derived from the locally learned representations. We also propose an online hyper-parameter tuning scheme which uses an online version of the REINFORCE algorithm to find a hyper-parameter distribution that maximizes the expected improvements in training loss. We show on several benchmarks that our two-part scheme of local representation matching and global adaptive hyper-parameters significantly improves performance and training robustness.
****
ROS-HPL: Robotic Object Search with Hierarchical Policy Learning and Intrinsic-Extrinsic Modeling
Xin Ye,Shibin Zheng,Yezhou Yang
Despite significant progress in Robotic Object Search (ROS) over the recent years with deep reinforcement learning based approaches, the sparsity issue in reward setting as well as the lack of interpretability of the previous ROS approaches leave much to be desired. We present a novel policy learning approach for ROS, based on a hierarchical and interpretable modeling with intrinsic/extrinsic reward setting, to tackle these two challenges. More specifically, we train the low-level policy by deliberating between an action that achieves an immediate sub-goal and the one that is better suited for achieving the final goal. We also introduce a new evaluation metric, namely the extrinsic reward, as a harmonic measure of the object search success rate and the average steps taken. Experiments conducted with multiple settings on the House3D environment validate and show that the intelligent agent, trained with our model, can achieve a better object search performance (higher success rate with lower average steps, measured by SPL: Success weighted by inverse Path Length). In addition, we conduct studies w.r.t. the parameter that controls the weighted overall reward from intrinsic and extrinsic components. The results suggest it is critical to devise a proper trade-off strategy to perform the object search well.
****
Knockoff-Inspired Feature Selection via Generative Models
Marco F. Duarte,Siwei Feng
We propose a feature selection algorithm for supervised learning inspired by the recently introduced

knockoff framework for variable selection in statistical regression. While variable selection in statistics aims
to distinguish between true and false predictors, feature selection in machine learning aims to reduce the
dimensionality of the data while preserving the performance of the learning method. The knockoff framework
has attracted significant interest due to its strong control of false discoveries while preserving predictive
power. In contrast to the original approach and later variants that assume a given probabilistic model for the
variables, our proposed approach relies on data-driven generative models that learn mappings from data
space to a parametric space that characterizes the probability distribution of the data. Our approach
requires only the availability of mappings from data space to a distribution in parametric space and from
parametric space to a distribution in data space; thus, it can be integrated with multiple popular generative
models from machine learning. We provide example knockoff designs using a variational autoencoder and
a Gaussian process latent variable model. We also propose a knockoff score metric for a softmax classifier
that accounts for the contribution of each feature and its knockoff during supervised learning. Experimental
results with multiple benchmark datasets for feature selection showcase the advantages of our knockoff
designs and the knockoff framework with respect to existing approaches.
****

MetaPix: Few-Shot Video Retargeting
Jessica Lee,Deva Ramanan,Rohit Girdhar
We address the task of unsupervised retargeting of human actions from one video to another. We consider the challenging setting where only a few frames of the target is available. The core of our approach is a conditional generative model that can transcode input skeletal poses (automatically extracted with an off-the-shelf pose estimator) to output target frames. However, it is challenging to build a universal transcoder because humans can appear wildly different due to clothing and background scene geometry. Instead, we learn to adapt – or personalize – a universal generator to the particular human and background in the target. To do so, we make use of meta-learning to discover effective strategies for on-the-fly personalization. One significant benefit of meta-learning is that the personalized transcoder naturally enforces temporal coherence across its generated frames; all frames contain consistent clothing and background geometry of the target. We experiment on in-the-wild internet videos and images and show our approach improves over widely-used baselines for the task.

****

SlowMo: Improving Communication-Efficient Distributed SGD with Slow Momentum
Jianyu Wang,Vinayak Tantia,Nicolas Ballas,Michael Rabbat
Distributed optimization is essential for training large models on large datasets. Multiple approaches have been proposed to reduce the communication overhead in distributed training, such as synchronizing only after performing multiple local SGD steps, and decentralized methods (e.g., using gossip algorithms) to decouple communications among workers. Although these methods run faster than AllReduce-based methods, which use blocking communication before every update, the resulting models may be less accurate after the same number of updates. Inspired by the BMUF method of Chen & Huo (2016), we propose a slow momentum (SlowMo) framework, where workers periodically synchronize and perform a momentum update, after multiple iterations of a base optimization algorithm. Experiments on image classification and machine translation tasks demonstrate that SlowMo consistently yields improvements in optimization and generalization performance relative to the

base optimizer, even when the additional overhead is amortized over many updates so that the SlowMo runtime is on par with that of the base optimizer. We provide theoretical convergence guarantees showing that SlowMo converges to a stationary point of smooth non-convex losses. Since BMUF can be expressed through the SlowMo framework, our results also correspond to the first theoretical convergence guarantees for BMUF.

****

Stochastic Prototype Embeddings

Tyler R. Scott,Karl Ridgeway,Michael C. Mozer

Supervised deep-embedding methods project inputs of a domain to a representational space in which same-class instances lie near one another and different-class instances lie far apart. We propose a probabilistic method that treats embeddings as random variables. Extending a state-of-the-art deterministic method, Prototypical Networks (Snell et al., 2017), our approach supposes the existence of a class prototype around which class instances are Gaussian distributed. The prototype posterior is a product distribution over labeled instances, and query instances are classified by marginalizing relative prototype proximity over embedding uncertainty. We describe an efficient sampler for approximate inference that allows us to train the model at roughly the same space and time cost as its deterministic sibling. Incorporating uncertainty improves performance on few-shot learning and gracefully handles label noise and out-of-distribution inputs. Compared to the state-of-the-art stochastic method, Hedged Instance Embeddings (Oh et al., 2019), we achieve superior large- and open-set classification accuracy. Our method also aligns class-discriminating features with the axes of the embedding space, yielding an interpretable, disentangled representation.

****

Way Off-Policy Batch Deep Reinforcement Learning of Human Preferences in Dialog

Natasha Jaques,Asma Ghandeharioun,Judy Hanwen Shen,Craig Ferguson,Agata Lapedriza,Noah Jones,Shixiang Gu,Rosalind Picard

Most deep reinforcement learning (RL) systems are not able to learn effectively from off-policy data, especially if they cannot explore online in the environment. This is a critical shortcoming for applying RL to real-world problems where collecting data is expensive, and models must be tested offline before being deployed to interact with the environment -- e.g. systems that learn from human interaction. Thus, we develop a novel class of off-policy batch RL algorithms which use KL-control to penalize divergence from a pre-trained prior model of probable actions. This KL-constraint reduces extrapolation error, enabling effective offline learning, without exploration, from a fixed batch of data. We also use dropout-based uncertainty estimates to lower bound the target Q-values as a more efficient alternative to Double Q-Learning. This Way Off-Policy (WOP) algorithm is tested on both traditional RL tasks from OpenAI Gym, and on the problem of open-domain dialog generation; a challenging reinforcement learning problem with a 20,000 dimensional action space. WOP allows for the extraction of multiple different reward functions post-hoc from collected human interaction data, and can learn effectively from all of these. We test real-world generalization by deploying dialog models live to converse with humans in an open-domain setting, and demonstrate that WOP achieves significant improvements over state-of-the-art prior methods in batch deep RL.

****

Targeted sampling of enlarged neighborhood via Monte Carlo tree search for TSP

Zhang-Hua Fu,Kai-Bin Qiu,Meng Qiu,Hongyuan Zha

The travelling salesman problem (TSP) is a well-known combinatorial optimization problem with a variety of real-life applications. We tackle TSP by incorporating machine learning methodology and leveraging the variable neighborhood search strategy. More precisely, the search process is considered as a Markov decision process (MDP), where a 2-opt local search is used to search within a small neighborhood, while a Monte Carlo tree search (MCTS) method (which iterates through simulation, selection and back-propagation steps), is used to sample a number of targeted actions within an enlarged neighborhood. This new paradigm clearly disti

nguishes itself from the existing machine learning (ML) based paradigms for solv ing the TSP, which either uses an end-to-end ML model, or simply applies traditi onal techniques after ML for post optimization. Experiments based on two public data sets show that, our approach clearly dominates all the existing learning ba sed TSP algorithms in terms of performance, demonstrating its high potential on the TSP. More importantly, as a general framework without complicated hand-craft ed rules, it can be readily extended to many other combinatorial optimization pr oblems.
****
Black-box Adversarial Attacks with Bayesian Optimization
Satya Narayan Shukla,Anit Kumar Sahu,Devin Willmott,J. Zico Kolter
We focus on the problem of black-box adversarial attacks, where the aim is to ge nerate adversarial examples using information limited to loss function evaluatio ns of input-output pairs. We use Bayesian optimization (BO) to specifically cater to scenarios involving low query budgets to develop query efficient advers arial attacks. We alleviate the issues surrounding BO in regards to optimizing h igh dimensional deep learning models by effective dimension upsampling technique s. Our proposed approach achieves performance comparable to the state of the art black-box adversarial attacks albeit with a much lower average query count. In particular, in low query budget regimes, our proposed method reduces the query c ount up to 80% with respect to the state of the art methods.
****
Pseudo-LiDAR++: Accurate Depth for 3D Object Detection in Autonomous Driving
Yurong You,Yan Wang,Wei-Lun Chao,Divyansh Garg,Geoff Pleiss,Bharath Hariharan,Ma rk Campbell,Kilian Q. Weinberger
Detecting objects such as cars and pedestrians in 3D plays an indispensable role in autonomous driving. Existing approaches largely rely on expensive LiDAR sens ors for accurate depth information. While recently pseudo-LiDAR has been introdu ced as a promising alternative, at a much lower cost based solely on stereo imag es, there is still a notable performance gap.
In this paper we provide substantial advances to the pseudo-LiDAR framework thro ugh improvements in stereo depth estimation. Concretely, we adapt the stereo net work architecture and loss function to be more aligned with accurate depth estim ation of faraway objects --- currently the primary weakness of pseudo-LiDAR. Fur ther, we explore the idea to leverage cheaper but extremely sparse LiDAR sensors , which alone provide insufficient information for 3D detection, to de-bias our depth estimation. We propose a depth-propagation algorithm, guided by the initia l depth estimates, to diffuse these few exact measurements across the entire dep th map. We show on the KITTI object detection benchmark that our combined approa ch yields substantial improvements in depth estimation and stereo-based 3D objec t detection --- outperforming the previous state-of-the-art detection accuracy f or faraway objects by 40%. Our code is available at https://github.com/mileyan/P seudo_Lidar_V2.
****
Learning to Combat Compounding-Error in Model-Based Reinforcement Learning
Chenjun Xiao,Yifan Wu,Chen Ma,Dale Schuurmans,Martin Müller
Despite its potential to improve sample complexity versus model-free approaches, model-based reinforcement learning can fail catastrophically if the model is in accurate. An algorithm should ideally be able to trust an imperfect model over a reasonably long planning horizon, and only rely on model-free updates when the model errors get infeasibly large. In this paper, we investigate techniques for choosing the planning horizon on a state-dependent basis, where a state's planni ng horizon is determined by the maximum cumulative model error around that state . We demonstrate that these state-dependent model errors can be learned with Tem poral Difference methods, based on a novel approach of temporally decomposing th e cumulative model errors. Experimental results show that the proposed method ca n successfully adapt the planning horizon to account for state-dependent model a ccuracy, significantly improving the efficiency of policy learning compared to model-based and model-free baselines.

****
Understanding Attention Mechanisms
Bingyuan Liu,Yogesh Balaji,Lingzhou Xue,Martin Renqiang Min
Attention mechanisms have advanced the state of the art in several machine learning tasks. Despite significant empirical gains, there is a lack of theoretical analyses on understanding their effectiveness. In this paper, we address this problem by studying the landscape of population and empirical loss functions of attention-based neural networks. Our results show that, under mild assumptions, every local minimum of a two-layer global attention model has low prediction error, and attention models require lower sample complexity than models not employing attention. We then extend our analyses to the popular self-attention model, proving that they deliver consistent predictions with a more expressive class of functions. Additionally, our theoretical results provide several guidelines for designing attention mechanisms. Our findings are validated with satisfactory experimental results on MNIST and IMDB reviews dataset.
****
Beyond GANs: Transforming without a Target Distribution
Matthew Amodio,David van Dijk,Ruth Montgomery,Guy Wolf,Smita Krishnaswamy
While generative neural networks can learn to transform a specific input dataset into a specific target dataset, they require having just such a paired set of input/output datasets. For instance, to fool the discriminator, a generative adversarial network (GAN) exclusively trained to transform images of black-haired *men* to blond-haired *men* would need to change gender-related characteristics as well as hair color when given images of black-haired *women* as input. This is problematic, as often it is possible to obtain *a* pair of (source, target) distributions but then have a second source distribution where the target distribution is unknown. The computational challenge is that generative models are good at generation within the manifold of the data that they are trained on. However, generating new samples outside of the manifold or extrapolating "out-of-sample" is a much harder problem that has been less well studied. To address this, we introduce a technique called *neuron editing* that learns how neurons encode an edit for a particular transformation in a latent space. We use an autoencoder to decompose the variation within the dataset into activations of different neurons and generate transformed data by defining an editing transformation on those neurons. By performing the transformation in a latent trained space, we encode fairly complex and non-linear transformations to the data with much simpler distribution shifts to the neuron's activations. Our technique is general and works on a wide variety of data domains and applications. We first demonstrate it on image transformations and then move to our two main biological applications: removal of batch artifacts representing unwanted noise and modeling the effect of drug treatments to predict synergy between drugs.
****
Four Things Everyone Should Know to Improve Batch Normalization
Cecilia Summers,Michael J. Dinneen
A key component of most neural network architectures is the use of normalization layers, such as Batch Normalization. Despite its common use and large utility in optimizing deep architectures, it has been challenging both to generically improve upon Batch Normalization and to understand the circumstances that lend themselves to other enhancements. In this paper, we identify four improvements to the generic form of Batch Normalization and the circumstances under which they work, yielding performance gains across all batch sizes while requiring no additional computation during training. These contributions include proposing a method for reasoning about the current example in inference normalization statistics, fixing a training vs. inference discrepancy; recognizing and validating the powerful regularization effect of Ghost Batch Normalization for small and medium batch sizes; examining the effect of weight decay regularization on the scaling and shifting parameters γ and β; and identifying a new normalization algorithm for very small batch sizes by combining the strengths of Batch and Group Normalization. We validate our results empirically on six datasets: CIFAR-100, SVHN, Caltech-256, Oxford Flowers-102, CUB-2011, and ImageNet.

****
Learning to solve the credit assignment problem
Benjamin James Lansdell,Prashanth Ravi Prakash,Konrad Paul Kording
Backpropagation is driving today's artificial neural networks (ANNs). However, despite extensive research, it remains unclear if the brain implements this algorithm. Among neuroscientists, reinforcement learning (RL) algorithms are often seen as a realistic alternative: neurons can randomly introduce change, and use unspecific feedback signals to observe their effect on the cost and thus approximate their gradient. However, the convergence rate of such learning scales poorly with the number of involved neurons. Here we propose a hybrid learning approach. Each neuron uses an RL-type strategy to learn how to approximate the gradients that backpropagation would provide. We provide proof that our approach converges to the true gradient for certain classes of networks. In both feedforward and convolutional networks, we empirically show that our approach learns to approximate the gradient, and can match the performance of gradient-based learning. Learning feedback weights provides a biologically plausible mechanism of achieving good performance, without the need for precise, pre-specified learning rules.
****
Improving Multi-Manifold GANs with a Learned Noise Prior
Matthew Amodio,Smita Krishnaswamy
Generative adversarial networks (GANs) learn to map samples from a noise distribution to a chosen data distribution. Recent work has demonstrated that GANs are consequently sensitive to, and limited by, the shape of the noise distribution. For example, a single generator struggles to map continuous noise (e.g. a uniform distribution) to discontinuous output (e.g. separate Gaussians) or complex output (e.g. intersecting parabolas). We address this problem by learning to generate from multiple models such that the generator's output is actually the combination of several distinct networks. We contribute a novel formulation of multi-generator models where we learn a prior over the generators conditioned on the noise, parameterized by a neural network. Thus, this network not only learns the optimal rate to sample from each generator but also optimally shapes the noise received by each generator. The resulting Noise Prior GAN (NPGAN) achieves expressivity and flexibility that surpasses both single generator models and previous multi-generator models.
****
Overparameterized Neural Networks Can Implement Associative Memory
Adityanarayanan Radhakrishnan,Mikhail Belkin,Caroline Uhler
Identifying computational mechanisms for memorization and retrieval is a long-standing problem at the intersection of machine learning and neuroscience. In this work, we demonstrate empirically that overparameterized deep neural networks trained using standard optimization methods provide a mechanism for memorization and retrieval of real-valued data. In particular, we show that overparameterized autoencoders store training examples as attractors, and thus, can be viewed as implementations of associative memory with the retrieval mechanism given by iterating the map. We study this phenomenon under a variety of common architectures and optimization methods and construct a network that can recall 500 real-valued images without any apparent spurious attractor states. Lastly, we demonstrate how the same mechanism allows encoding sequences, including movies and audio, instead of individual examples. Interestingly, this appears to provide an even more efficient mechanism for storage and retrieval than autoencoding single instances.

****
Bayesian Residual Policy Optimization: Scalable Bayesian Reinforcement Learning with Clairvoyant Experts
Gilwoo Lee,Brian Hou,Sanjiban Choudhury,Siddhartha S. Srinivasa
Informed and robust decision making in the face of uncertainty is critical for robots that perform physical tasks alongside people. We formulate this as a Bayesian Reinforcement Learning problem over latent Markov Decision Processes (MDPs). While Bayes-optimality is theoretically the gold standard, existing algorithms

do not scale well to continuous state and action spaces. We propose a scalable s
olution that builds on the following insight: in the absence of uncertainty, eac
h latent MDP is easier to solve. We split the challenge into two simpler compone
nts. First, we obtain an ensemble of clairvoyant experts and fuse their advice t
o compute a baseline policy. Second, we train a Bayesian residual policy to impr
ove upon the ensemble's recommendation and learn to reduce uncertainty. Our algo
rithm, Bayesian Residual Policy Optimization (BRPO), imports the scalability of
policy gradient methods as well as the initialization from prior models. BRPO si
gnificantly improves the ensemble of experts and drastically outperforms existin
g adaptive RL methods.
****

Sampling-Free Learning of Bayesian Quantized Neural Networks
Jiahao Su,Milan Cvitkovic,Furong Huang
Bayesian learning of model parameters in neural networks is important in scenari
os where estimates with well-calibrated uncertainty are important. In this paper
, we propose Bayesian quantized networks (BQNs), quantized neural networks (QNNs
) for which we learn a posterior distribution over their discrete parameters. We
 provide a set of efficient algorithms for learning and prediction in BQNs witho
ut the need to sample from their parameters or activations, which not only allow
s for differentiable learning in quantized models but also reduces the variance
in gradients estimation. We evaluate BQNs on MNIST, Fashion-MNIST and KMNIST cla
ssification datasets compared against bootstrap ensemble of QNNs (E-QNN). We dem
onstrate BQNs achieve both lower predictive errors and better-calibrated uncerta
inties than E-QNN (with less than 20% of the negative log-likelihood).
****

A Hierarchy of Graph Neural Networks Based on Learnable Local Features
Michael Lingzhi Li,Meng Dong,Jiawei Zhou,Alexander M. Rush
Graph neural networks (GNNs) are a powerful tool to learn representations on gra
phs by iteratively aggregating features from node neighbourhoods. Many variant m
odels have been proposed, but there is limited understanding on both how to comp
are different architectures and how to construct GNNs systematically. Here, we p
ropose a hierarchy of GNNs based on their aggregation regions. We derive theoret
ical results about the discriminative power and feature representation capabilit
ies of each class. Then, we show how this framework can be utilized to systemati
cally construct arbitrarily powerful GNNs. As an example, we construct a simple
architecture that exceeds the expressiveness of the Weisfeiler-Lehman graph isom
orphism test. We empirically validate our theory on both synthetic and real-worl
d benchmarks, and demonstrate our example's theoretical power translates to stat
e-of-the-art results on node classification, graph classification, and graph reg
ression tasks.
****

DeFINE: Deep Factorized Input Token Embeddings for Neural Sequence Modeling
Sachin Mehta,Rik Koncel-Kedziorski,Mohammad Rastegari,Hannaneh Hajishirzi
For sequence models with large vocabularies, a majority of network parameters li
e in the input and output layers. In this work, we describe a new method, DeFINE
, for learning deep token representations efficiently. Our architecture uses a h
ierarchical structure with novel skip-connections which allows for the use of lo
w dimensional input and output layers, reducing total parameters and training ti
me while delivering similar or better performance versus existing methods. DeFIN
E can be incorporated easily in new or existing sequence models. Compared to sta
te-of-the-art methods including adaptive input representations, this technique r
esults in a 6% to 20% drop in perplexity. On WikiText-103, DeFINE reduces the to
tal parameters of Transformer-XL by half with minimal impact on performance. On
the Penn Treebank, DeFINE improves AWD-LSTM by 4 points with a 17% reduction in
parameters, achieving comparable performance to state-of-the-art methods with fe
wer parameters. For machine translation, DeFINE improves the efficiency of the T
ransformer model by about 1.4 times while delivering similar performance.
****

NEURAL EXECUTION ENGINES
Yujun Yan,Kevin Swersky,Danai Koutra,Parthasarathy Ranganathan,Milad Hashemi

Turing complete computation and reasoning are often regarded as necessary pre- c
ursors to general intelligence. There has been a significant body of work studyi
ng neural networks that mimic general computation, but these networks fail to ge
neralize to data distributions that are outside of their training set. We study
this problem through the lens of fundamental computer science problems: sorting
and graph processing. We modify the masking mechanism of a transformer in order
to allow them to implement rudimentary functions with strong generalization. We
call this model the Neural Execution Engine, and show that it learns, through su
pervision, to numerically compute the basic subroutines comprising these algorit
hms with near perfect accuracy. Moreover, it retains this level of accuracy whil
e generalizing to unseen data and long sequences outside of the training distrib
ution.
****

Learning to Make Generalizable and Diverse Predictions for Retrosynthesis
Benson Chen,Tianxiao Shen,Tommi S. Jaakkola,Regina Barzilay
We propose a new model for making generalizable and diverse retrosynthetic react
ion predictions. Given a target compound, the task is to predict the likely chem
ical reactants to produce the target. This generative task can be framed as a se
quence-to-sequence problem by using the SMILES representations of the molecules.
 Building on top of the popular Transformer architecture, we propose two novel p
re-training methods that construct relevant auxiliary tasks (plausible reactions
) for our problem. Furthermore, we incorporate a discrete latent variable model
into the architecture to encourage the model to produce a diverse set of alterna
tive predictions. On the 50k subset of reaction examples from the United States
patent literature (USPTO-50k) benchmark dataset, our model greatly improves perf
ormance over the baseline, while also generating predictions that are more diver
se.
****

Disentangled GANs for Controllable Generation of High-Resolution Images
Weili Nie,Tero Karras,Animesh Garg,Shoubhik Debhath,Anjul Patney,Ankit B. Patel,
Anima Anandkumar
Generative adversarial networks (GANs) have achieved great success at generating
 realistic samples. However, achieving disentangled and controllable generation
still remains challenging for GANs, especially in the high-resolution image doma
in. Motivated by this, we introduce AC-StyleGAN, a combination of AC-GAN and Sty
leGAN, for demonstrating that the controllable generation of high-resolution ima
ges is possible with sufficient supervision. More importantly, only using 5% of
the labelled data significantly improves the disentanglement quality. Inspired b
y the observed separation of fine and coarse styles in StyleGAN, we then extend
AC-StyleGAN to a new image-to-image model called FC-StyleGAN for semantic manipu
lation of fine-grained factors in a high-resolution image. In experiments, we sh
ow that FC-StyleGAN performs well in only controlling fine-grained factors, with
 the use of instance normalization, and also demonstrate its good generalization
 ability to unseen images. Finally, we create two new datasets -- Falcor3D and I
saac3D with higher resolution, more photorealism, and richer variation, as compa
red to existing disentanglement datasets.
****

Continuous Graph Flow
Zhiwei Deng,Megha Nawhal,Lili Meng,Greg Mori
In this paper, we propose Continuous Graph Flow, a generative continuous flow ba
sed method that aims to model complex distributions of graph-structured data. O
nce learned, the model can be applied to an arbitrary graph, defining a probabil
ity density over the random variables represented by the graph. It is formulated
 as an ordinary differential equation system with shared and reusable functions
that operate over the graphs. This leads to a new type of neural graph message
passing scheme that performs continuous message passing over time. This class of
 models offers several advantages: a flexible representation that can generalize
 to variable data dimensions; ability to model dependencies in complex data dist
ributions; reversible and memory-efficient; and exact and efficient computation
of the likelihood of the data. We demonstrate the effectiveness of our model on

a diverse set of generation tasks across different domains: graph generation, im age puzzle generation, and layout generation from scene graphs. Our proposed mod el achieves significantly better performance compared to state-of-the-art model s.
****

Wasserstein-Bounded Generative Adversarial Networks
Peng Zhou,Bingbing Ni,Lingxi Xie,Xiaopeng Zhang,Hang Wang,Cong Geng,Qi Tian
In the field of Generative Adversarial Networks (GANs), how to design a stable t raining strategy remains an open problem. Wasserstein GANs have largely promoted the stability over the original GANs by introducing Wasserstein distance, but s till remain unstable and are prone to a variety of failure modes. In this paper, we present a general framework named Wasserstein-Bounded GAN (WBGAN), which imp roves a large family of WGAN-based approaches by simply adding an upper-bound co nstraint to the Wasserstein term. Furthermore, we show that WBGAN can reasonably measure the difference of distributions which almost have no intersection. Expe riments demonstrate that WBGAN can stabilize as well as accelerate convergence i n the training processes of a series of WGAN-based variants.
****

DBA: Distributed Backdoor Attacks against Federated Learning
Chulin Xie,Keli Huang,Pin-Yu Chen,Bo Li
Backdoor attacks aim to manipulate a subset of training data by injecting advers arial triggers such that machine learning models trained on the tampered dataset will make arbitrarily (targeted) incorrect prediction on the testset with the s ame trigger embedded. While federated learning (FL) is capable of aggregating in formation provided by different parties for training a better model, its distrib uted learning methodology and inherently heterogeneous data distribution across parties may bring new vulnerabilities. In addition to recent centralized backdoo r attacks on FL where each party embeds the same global trigger during training, we propose the distributed backdoor attack (DBA) --- a novel threat assessment framework developed by fully exploiting the distributed nature of FL. DBA decomp oses a global trigger pattern into separate local patterns and embed them into t he training set of different adversarial parties respectively. Compared to stand ard centralized backdoors, we show that DBA is substantially more persistent and stealthy against FL on diverse datasets such as finance and image data. We cond uct extensive experiments to show that the attack success rate of DBA is signifi cantly higher than centralized backdoors under different settings. Moreover, we find that distributed attacks are indeed more insidious, as DBA can evade two st ate-of-the-art robust FL algorithms against centralized backdoors. We also provi de explanations for the effectiveness of DBA via feature visual interpretation a nd feature importance ranking.
To further explore the properties of DBA, we test the attack performance by vary ing different trigger factors, including local trigger variations (size, gap, an d location), scaling factor in FL, data distribution, and poison ratio and inter val. Our proposed DBA and thorough evaluation results shed lights on characteriz ing the robustness of FL.
****

Learning Generative Models using Denoising Density Estimators
Siavash Bigdeli,Geng Lin,Tiziano Portenier,Andrea Dunbar,Matthias Zwicker
Learning generative probabilistic models that can estimate the continuous densit y given a set of samples, and that can sample from that density is one of the fu ndamental challenges in unsupervised machine learning. In this paper we introduc e a new approach to obtain such models based on what we call denoising density e stimators (DDEs). A DDE is a scalar function, parameterized by a neural network, that is efficiently trained to represent a kernel density estimator of the data . In addition, we show how to leverage DDEs to develop a novel approach to obtai n generative models that sample from given densities. We prove that our algorith ms to obtain both DDEs and generative models are guaranteed to converge to the c orrect solutions. Advantages of our approach include that we do not require spec ific network architectures like in normalizing flows, ODE solvers as in continuo us normalizing flows, nor do we require adversarial training as in generative ad

versarial networks (GANs). Finally, we provide experimental results that demonst
rate practical applications of our technique.

****
Fast is better than free: Revisiting adversarial training
Eric Wong,Leslie Rice,J. Zico Kolter
Adversarial training, a method for learning robust deep networks, is typically a
ssumed to be more expensive than traditional training due to the necessity of co
nstructing adversarial examples via a first-order method like projected gradient
 decent (PGD). In this paper, we make the surprising discovery that it is possi
ble to train empirically robust models using a much weaker and cheaper adversary
, an approach that was previously believed to be ineffective, rendering the meth
od no more costly than standard training in practice. Specifically, we show tha
t adversarial training with the fast gradient sign method (FGSM), when combined
 with random initialization, is as effective as PGD-based training but has signif
icantly lower cost. Furthermore we show that FGSM adversarial training can be f
urther accelerated by using standard techniques for efficient training of deep n
etworks, allowing us to learn a robust CIFAR10 classifier with 45% robust accura
cy at epsilon=8/255 in 6 minutes, and a robust ImageNet classifier with 43% robu
st accuracy at epsilon=2/255 in 12 hours, in comparison to past work based on ``
free'' adversarial training which took 10 and 50 hours to reach the same respect
ive thresholds.
****
LOSSLESS SINGLE IMAGE SUPER RESOLUTION FROM LOW-QUALITY JPG IMAGES
Yong Shi,Biao Li,Bo Wang,Zhiquan Qi,Jiabin Liu,Fan Meng
Super Resolution (SR) is a fundamental and important low-level computer vision (
CV) task. Different from traditional SR models, this study concentrates on a spe
cific but realistic SR issue: How can we obtain satisfied SR results from compre
ssed JPG (C-JPG) image, which widely exists on the Internet. In general, C-JPG c
an release storage space while keeping considerable quality in visual. However,
further image processing operations, e.g., SR, will suffer from enlarging inner
artificial details and result in unacceptable outputs. To address this problem,
we propose a novel SR structure with two specifically designed components, as we
ll as a cycle loss. In short, there are mainly three contributions to this paper
. First, our research can generate high-qualified SR images for prevalent C-JPG
images. Second, we propose a functional sub-model to recover information for C-J
PG images, instead of the perspective of noise elimination in traditional SR app
roaches. Third, we further integrate cycle loss into SR solver to build a hybrid
 loss function for better SR generation. Experiments show that our approach achi
eves outstanding performance among state-of-the-art methods.
****
iWGAN: an Autoencoder WGAN for Inference
Yao Chen,Qingyi Gao,Xiao Wang
Generative Adversarial Networks (GANs) have been impactful on many problems and
applications but suffer from unstable training. Wasserstein GAN (WGAN) leverages
 the Wasserstein distance to avoid the caveats in the minmax two-player training
 of GANs but has other defects such as mode collapse and lack of metric to detec
t the convergence. We introduce a novel inference WGAN (iWGAN) model, which is a
 principled framework to fuse auto-encoders and WGANs. The iWGAN jointly learns
an encoder network and a generative network using an iterative primal dual optim
ization process. We establish the generalization error bound of iWGANs. We furth
er provide a rigorous probabilistic interpretation of our model under the framew
ork of maximum likelihood estimation. The iWGAN, with a clear stopping criteria,
 has many advantages over other autoencoder GANs. The empirical experiments show
 that our model greatly mitigates the symptom of mode collapse, speeds up the co
nvergence, and is able to provide a measurement of quality check for each indivi
dual sample. We illustrate the ability of iWGANs by obtaining a competitive and
stable performance with state-of-the-art for benchmark datasets.
****
BERT-AL: BERT for Arbitrarily Long Document Understanding

Ruixuan Zhang,Zhuoyu Wei,Yu Shi,Yining Chen
Pretrained language models attract lots of attentions, and they take advantage of the two-stages training process: pretraining on huge corpus and finetuning on specific tasks. Thereinto, BERT (Devlin et al., 2019) is a Transformer (Vaswani et al., 2017) based model and has been the state-of-the-art for many kinds of Nature Language Processing (NLP) tasks. However, BERT cannot take text longer than the maximum length as input since the maximum length is predefined during pretraining. When we apply BERT to long text tasks, e.g., document-level text summarization: 1) Truncating inputs by the maximum sequence length will decrease performance, since the model cannot capture long dependency and global information ranging the whole document. 2) Extending the maximum length requires re-pretraining which will cost a mass of time and computing resources. What's even worse is that the computational complexity will increase quadratically with the length, which will result in an unacceptable training time. To resolve these problems, we propose to apply Transformer to only model local dependency and recurrently capture long dependency by inserting multi-channel LSTM into each layer of BERT. The proposed model is named as BERT-AL (BERT for Arbitrarily Long Document Understanding) and it can accept arbitrarily long input without re-pretraining from scratch. We demonstrate BERT-AL's effectiveness on text summarization by conducting experiments on the CNN/Daily Mail dataset. Furthermore, our method can be adapted to other Transformer based models, e.g., XLNet (Yang et al., 2019) and RoBERTa (Liu et al., 2019), for various NLP tasks with long text.
****
Novelty Search in representational space for sample efficient exploration
Ruo Yu Tao,Vincent François-Lavet,Joelle Pineau
We present a new approach for efficient exploration which leverages a low-dimensional encoding of the environment learned with a combination of model-based and model-free objectives. Our approach uses intrinsic rewards that are based on a weighted distance of nearest neighbors in the low dimensional representational space to gauge novelty.
We then leverage these intrinsic rewards for sample-efficient exploration with planning routines in representational space.
One key element of our approach is that we perform more gradient steps in-between every environment step in order to ensure the model accuracy. We test our approach on a number of maze tasks, as well as a control problem and show that our exploration approach is more sample-efficient compared to strong baselines.
****
Switched linear projections and inactive state sensitivity for deep neural network interpretability
Lech Szymanski,Brendan McCane,Craig Atkinson
We introduce switched linear projections for expressing the activity of a neuron in a ReLU-based deep neural network in terms of a single linear projection in the input space. The method works by isolating the active subnetwork, a series of linear transformations, that completely determine the entire computation of the deep network for a given input instance. We also propose that for interpretability it is more instructive and meaningful to focus on the patterns that deactive the neurons in the network, which are ignored by the exisiting methods that implicitly track only the active aspect of the network's computation. We introduce a novel interpretability method for the inactive state sensitivity (Insens). Comparison against existing methods shows that Insens is more robust (in the presence of noise), more complete (in terms of patterns that affect the computation) and a very effective interpretability method for deep neural networks
****
An Optimization Principle Of Deep Learning?
Cheng Chen,Junjie Yang,Yi Zhou
Training deep neural networks (DNNs) has achieved great success in recent years. Modern DNN trainings utilize various types of training techniques that are developed in different aspects, e.g., activation functions for neurons, batch normalization for hidden layers, skip connections for network architecture and stochastic algorithms for optimization. Despite the effectiveness of these techniques,

it is still mysterious how they help accelerate DNN trainings in practice. In th
is paper, we propose an optimization principle that is parameterized by $\gamma>
0$ for stochastic algorithms in nonconvex and over-parameterized optimization. T
he principle guarantees the convergence of stochastic algorithms to a global min
imum with a monotonically diminishing parameter distance to the minimizer and le
ads to a $\mathcal{O}(1/\gamma K)$ sub-linear convergence rate, where K is the
 number of iterations. Through extensive experiments, we show that DNN trainings
 consistently obey the γ-optimization principle and its theoretical impli
cations. In particular, we observe that the trainings that apply the training te
chniques achieve accelerated convergence and obey the principle with a large $\g
amma$, which is consistent with the $\mathcal{O}(1/\gamma K)$ convergence rate r
esult under the optimization principle. We think the γ-optimization princ
iple captures and quantifies the impacts of various DNN training techniques and
can be of independent interest from a theoretical perspective.
****

Thieves on Sesame Street! Model Extraction of BERT-based APIs
Kalpesh Krishna,Gaurav Singh Tomar,Ankur P. Parikh,Nicolas Papernot,Mohit Iyyer
We study the problem of model extraction in natural language processing, in whic
h an adversary with only query access to a victim model attempts to reconstruct
a local copy of that model. Assuming that both the adversary and victim model fi
ne-tune a large pretrained language model such as BERT (Devlin et al., 2019), we
 show that the adversary does not need any real training data to successfully mo
unt the attack. In fact, the attacker need not even use grammatical or semantica
lly meaningful queries: we show that random sequences of words coupled with task
-specific heuristics form effective queries for model extraction on a diverse se
t of NLP tasks, including natural language inference and question answering. Our
 work thus highlights an exploit only made feasible by the shift towards transfe
r learning methods within the NLP community: for a query budget of a few hundred
 dollars, an attacker can extract a model that performs only slightly worse than
 the victim model. Finally, we study two defense strategies against model extrac
tion—membership classification and API watermarking—which while successful again
st some adversaries can also be circumvented by more clever ones.
****

Understanding Knowledge Distillation in Non-autoregressive Machine Translation
Chunting Zhou,Jiatao Gu,Graham Neubig
Non-autoregressive machine translation (NAT) systems predict a sequence of outpu
t tokens in parallel, achieving substantial improvements in generation speed com
pared to autoregressive models. Existing NAT models usually rely on the techniqu
e of knowledge distillation, which creates the training data from a pretrained a
utoregressive model for better performance. Knowledge distillation is empiricall
y useful, leading to large gains in accuracy for NAT models, but the reason for
this success has, as of yet, been unclear. In this paper, we first design system
atic experiments to investigate why knowledge distillation is crucial to NAT tra
ining. We find that knowledge distillation can reduce the complexity of data set
s and help NAT to model the variations in the output data. Furthermore, a strong
 correlation is observed between the capacity of an NAT model and the optimal co
mplexity of the distilled data for the best translation quality. Based on these
findings, we further propose several approaches that can alter the complexity of
 data sets to improve the performance of NAT models. We achieve the state-of-the
-art performance for the NAT-based models, and close the gap with the autoregres
sive baseline on WMT14 En-De benchmark.
****

Coordinated Exploration via Intrinsic Rewards for Multi-Agent Reinforcement Lear
ning
Shariq Iqbal,Fei Sha
Solving tasks with sparse rewards is one of the most important challenges in rei
nforcement learning. In the single-agent setting, this challenge has been addres
sed by introducing intrinsic rewards that motivate agents to explore unseen regi
ons of their state spaces. Applying these techniques naively to the multi-agent
setting results in agents exploring independently, without any coordination amon

g themselves. We argue that learning in cooperative multi-agent settings can be accelerated and improved if agents coordinate with respect to what they have explored. In this paper we propose an approach for learning how to dynamically select between different types of intrinsic rewards which consider not just what an individual agent has explored, but all agents, such that the agents can coordinate their exploration and maximize extrinsic returns. Concretely, we formulate the approach as a hierarchical policy where a high-level controller selects among sets of policies trained on different types of intrinsic rewards and the low-level controllers learn the action policies of all agents under these specific rewards. We demonstrate the effectiveness of the proposed approach in a multi-agent gridworld domain with sparse rewards, and then show that our method scales up to more complex settings by evaluating on the VizDoom platform.
****

Generative Teaching Networks: Accelerating Neural Architecture Search by Learning to Generate Synthetic Training Data
Felipe Petroski Such,Aditya Rawal,Joel Lehman,Kenneth Stanley,Jeff Clune
This paper investigates the intriguing question of whether we can create learning algorithms that automatically generate training data, learning environments, and curricula in order to help AI agents rapidly learn. We show that such algorithms are possible via Generative Teaching Networks (GTNs), a general approach that is applicable to supervised, unsupervised, and reinforcement learning. GTNs are deep neural networks that generate data and/or training environments that a learner (e.g.\ a freshly initialized neural network) trains on before being tested on a target task. We then differentiate \emph{through the entire learning process} via meta-gradients to update the GTN parameters to improve performance on the target task. GTNs have the beneficial property that they can theoretically generate any type of data or training environment, making their potential impact large. This paper introduces GTNs, discusses their potential, and showcases that they can substantially accelerate learning. We also demonstrate a practical and exciting application of GTNs: accelerating the evaluation of candidate architectures for neural architecture search (NAS), which is rate-limited by such evaluations, enabling massive speed-ups in NAS. GTN-NAS improves the NAS state of the art, finding higher performing architectures when controlling for the search proposal mechanism. GTN-NAS also is competitive with the overall state of the art approaches, which achieve top performance while using orders of magnitude less computation than typical NAS methods. Overall, GTNs represent a first step toward the ambitious goal of algorithms that generate their own training data and, in doing so, open a variety of interesting new research questions and directions.
****

Locality and Compositionality in Zero-Shot Learning
Tristan Sylvain,Linda Petrini,Devon Hjelm
In this work we study locality and compositionality in the context of learning representations for Zero Shot Learning (ZSL).
In order to well-isolate the importance of these properties in learned representations, we impose the additional constraint that, differently from most recent work in ZSL, no pre-training on different datasets (e.g. ImageNet) is performed.
The results of our experiment show how locality, in terms of small parts of the input, and compositionality, i.e. how well can the learned representations be expressed as a function of a smaller vocabulary, are both deeply related to generalization and motivate the focus on more local-aware models in future research directions for representation learning.
****

Optimistic Adaptive Acceleration for Optimization
Jun-Kun Wang,Xiaoyun Li,Ping Li
This paper considers a new variant of AMSGrad called Optimistic-AMSGrad. AMSGrad is a popular adaptive gradient based optimization algorithm that is widely used in training deep neural networks. The new variant assumes that mini-batch gradients in consecutive iterations have some underlying structure, which makes the gradients sequentially predictable. By exploiting the predictability and some ideas from Optimistic Online learning, the proposed algorithm can accelerate the co

nvergence and also enjoys a tighter regret bound. We evaluate Optimistic-AMSGrad and AMSGrad in terms of various performance measures (i.e., training loss, testing loss, and classification accuracy on training/testing data), which demonstrate that Optimistic-AMSGrad improves AMSGrad.
****

Situating Sentence Embedders with Nearest Neighbor Overlap
Lucy H. Lin,Noah A. Smith
As distributed approaches to natural language semantics have developed and diversified, embedders for linguistic units larger than words (e.g., sentences) have come to play an increasingly important role. To date, such embedders have been evaluated using benchmark tasks (e.g., GLUE) and linguistic probes. We propose a comparative approach, nearest neighbor overlap (N2O), that quantifies similarity between embedders in a task-agnostic manner. N2O requires only a collection of examples and is simple to understand: two embedders are more similar if, for the same set of inputs, there is greater overlap between the inputs' nearest neighbors. We use N2O to compare 21 sentence embedders and show the effects of different design choices and architectures.
****

Generalized Clustering by Learning to Optimize Expected Normalized Cuts
Azade Nazi,Will Hang,Anna Goldie,Sujith Ravi,Azalia Mirhoseini
We introduce a novel end-to-end approach for learning to cluster in the absence of labeled examples. Our clustering objective is based on optimizing normalized cuts, a criterion which measures both intra-cluster similarity as well as inter-cluster dissimilarity. We define a differentiable loss function equivalent to the expected normalized cuts. Unlike much of the work in unsupervised deep learning, our trained model directly outputs final cluster assignments, rather than embeddings that need further processing to be usable. Our approach generalizes to unseen datasets across a wide variety of domains, including text, and image. Specifically, we achieve state-of-the-art results on popular unsupervised clustering benchmarks (e.g., MNIST, Reuters, CIFAR-10, and CIFAR-100), outperforming the strongest baselines by up to 10.9%. Our generalization results are superior (by up to 21.9%) to the recent top-performing clustering approach with the ability to generalize.
****

Recurrent neural circuits for contour detection
Drew Linsley*,Junkyung Kim*,Alekh Ashok,Thomas Serre
We introduce a deep recurrent neural network architecture that approximates visual cortical circuits (Mély et al., 2018). We show that this architecture, which we refer to as the ■-net, learns to solve contour detection tasks with better sample efficiency than state-of-the-art feedforward networks, while also exhibiting a classic perceptual illusion, known as the orientation-tilt illusion. Correcting this illusion significantly reduces \gnetw contour detection accuracy by driving it to prefer low-level edges over high-level object boundary contours. Overall, our study suggests that the orientation-tilt illusion is a byproduct of neural circuits that help biological visual systems achieve robust and efficient contour detection, and that incorporating these circuits in artificial neural networks can improve computer vision.
****

Disentangling neural mechanisms for perceptual grouping
Junkyung Kim*,Drew Linsley*,Kalpit Thakkar,Thomas Serre
Forming perceptual groups and individuating objects in visual scenes is an essential step towards visual intelligence. This ability is thought to arise in the brain from computations implemented by bottom-up, horizontal, and top-down connections between neurons. However, the relative contributions of these connections to perceptual grouping are poorly understood. We address this question by systematically evaluating neural network architectures featuring combinations bottom-up, horizontal, and top-down connections on two synthetic visual tasks, which stress low-level "Gestalt" vs. high-level object cues for perceptual grouping. We show that increasing the difficulty of either task strains learning for networks that rely solely on bottom-up connections. Horizontal connections resolve strain

ing on tasks with Gestalt cues by supporting incremental grouping, whereas top-down connections rescue learning on tasks with high-level object cues by modifying coarse predictions about the position of the target object. Our findings dissociate the computational roles of bottom-up, horizontal and top-down connectivity, and demonstrate how a model featuring all of these interactions can more flexibly learn to form perceptual groups.
****
Adversarial Imitation Attack
Mingyi Zhou,Jing Wu,Yipeng Liu,Xiaolin Huang,Shuaicheng Liu,Liaqat Ali,Xiang Zhang,Ce Zhu
Deep learning models are known to be vulnerable to adversarial examples. A practical adversarial attack should require as little as possible knowledge of attacked models T. Current substitute attacks need pre-trained models to generate adversarial examples and their attack success rates heavily rely on the transferability of adversarial examples. Current score-based and decision-based attacks require lots of queries for the T. In this study, we propose a novel adversarial imitation attack. First, it produces a replica of the T by a two-player game like the generative adversarial networks (GANs). The objective of the generative model G is to generate examples which lead D returning different outputs with T. The objective of the discriminative model D is to output the same labels with T under the same inputs. Then, the adversarial examples generated by D are utilized to fool the T. Compared with the current substitute attacks, imitation attack can use less training data to produce a replica of T and improve the transferability of adversarial examples. Experiments demonstrate that our imitation attack requires less training data than the black-box substitute attacks, but achieves an attack success rate close to the white-box attack on unseen data with no query.
****
Regularizing Trajectories to Mitigate Catastrophic Forgetting
Paul Michel,Elisabeth Salesky,Graham Neubig
Regularization-based continual learning approaches generally prevent catastrophic forgetting by augmenting the training loss with an auxiliary objective. However in most practical optimization scenarios with noisy data and/or gradients, it is possible that stochastic gradient descent can inadvertently change critical parameters.
In this paper, we argue for the importance of regularizing optimization trajectories directly. We derive a new co-natural gradient update rule for continual learning whereby the new task gradients are preconditioned with the empirical Fisher information of previously learnt tasks. We show that using the co-natural gradient systematically reduces forgetting in continual learning. Moreover, it helps combat overfitting when learning a new task in a low resource scenario.
****
Semantic Pruning for Single Class Interpretability
Kamila Abdiyeva,Martin Lukac,Kanat Alimanov
Convolutional Neural Networks (CNN) have achieved state-of-the-art performance in different computer vision tasks, but at a price of being computationally and power intensive. At the same time, only a few attempts were made toward a deeper understanding of CNNs. In this work, we propose to use semantic pruning technique toward not only CNN optimization but also as a way toward getting some insight information on convolutional filters correlation and interference. We start with a pre-trained network and prune it until it behaves as a single class classifier for a selected class. Unlike the more traditional approaches which apply retraining to the pruned CNN, the proposed semantic pruning does not use retraining. Conducted experiments showed that a) for each class there is a pruning ration which allows removing filters with either an increase or no loss of classification accuracy, b) pruning can improve the interference between filters used for classification of different classes c) effect between classification accuracy and correlation between pruned filters groups specific for different classes.
****
Analyzing the Role of Model Uncertainty for Electronic Health Records
Michael W. Dusenberry,Dustin Tran,Edward Choi,Jonas Kemp,Jeremy Nixon,Ghassen Je

rfel,Katherine Heller,Andrew M. Dai
In medicine, both ethical and monetary costs of incorrect predictions can be sig
nificant, and the complexity of the problems often necessitates increasingly com
plex models. Recent work has shown that changing just the random seed is enough
for otherwise well-tuned deep neural networks to vary in their individual predic
ted probabilities. In light of this, we investigate the role of model uncertaint
y methods in the medical domain. Using RNN ensembles and various Bayesian RNNs,
we show that population-level metrics, such as AUC-PR, AUC-ROC, log-likelihood,
and calibration error, do not capture model uncertainty. Meanwhile, the presence
 of significant variability in patient-specific predictions and optimal decision
s motivates the need for capturing model uncertainty. Understanding the uncertai
nty for individual patients is an area with clear clinical impact, such as deter
mining when a model decision is likely to be brittle. We further show that RNNs
with only Bayesian embeddings can be a more efficient way to capture model uncer
tainty compared to ensembles, and we analyze how model uncertainty is impacted a
cross individual input features and patient subgroups.
****

Chameleon: Adaptive Code Optimization for Expedited Deep Neural Network Compilat
ion
Byung Hoon Ahn,Prannoy Pilligundla,Amir Yazdanbakhsh,Hadi Esmaeilzadeh
Achieving faster execution with shorter compilation time can foster further dive
rsity and innovation in neural networks. However, the current paradigm of execut
ing neural networks either relies on hand-optimized libraries, traditional compi
lation heuristics, or very recently genetic algorithms and other stochastic meth
ods. These methods suffer from frequent costly hardware measurements rendering t
hem not only too time consuming but also suboptimal. As such, we devise a soluti
on that can learn to quickly adapt to a previously unseen design space for code
optimization, both accelerating the search and improving the output performance.
 This solution dubbed Chameleon leverages reinforcement learning whose solution
takes fewer steps to converge, and develops an adaptive sampling algorithm that
not only focuses on the costly samples (real hardware measurements) on represent
ative points but also uses a domain-knowledge inspired logic to improve the samp
les itself. Experimentation with real hardware shows that Chameleon provides 4.4
5x speed up in optimization time over AutoTVM, while also improving inference ti
me of the modern deep networks by 5.6%.
****

Weakly-supervised Knowledge Graph Alignment with Adversarial Learning
Meng Qu,Jian Tang,Yoshua Bengio
This paper studies aligning knowledge graphs from different sources or languages
. Most existing methods train supervised methods for the alignment, which usuall
y require a large number of aligned knowledge triplets. However, such a large nu
mber of aligned knowledge triplets may not be available or are expensive to obta
in in many domains. Therefore, in this paper we propose to study aligning knowle
dge graphs in fully-unsupervised or weakly-supervised fashion, i.e., without or
with only a few aligned triplets. We propose an unsupervised framework to align
the entity and relation embddings of different knowledge graphs with an adversar
ial learning framework. Moreover, a regularization term which maximizes the mutu
al information between the embeddings of different knowledge graphs is used to m
itigate the problem of mode collapse when learning the alignment functions. Such
 a framework can be further seamlessly integrated with existing supervised metho
ds by utilizing a limited number of aligned triples as guidance. Experimental re
sults on multiple datasets prove the effectiveness of our proposed approach in b
oth the unsupervised and the weakly-supervised settings.
****

Auto Completion of User Interface Layout Design Using Transformer-Based Tree Dec
oders
Yang Li,Julien Amelot,Xin Zhou,Samy Bengio,Si Si
It has been of increasing interest in the field to develop automatic machineries
 to facilitate the design process. In this paper, we focus on assisting graphica
l user interface (UI) layout design, a crucial task in app development. Given a

partial layout, which a designer has entered, our model learns to complete the l
ayout by predicting the remaining UI elements with a correct position and dimens
ion as well as the hierarchical structures. Such automation will significantly e
ase the effort of UI designers and developers. While we focus on interface layou
t prediction, our model can be generally applicable for other layout prediction
problems that involve tree structures and 2-dimensional placements. Particularly
, we design two versions of Transformer-based tree decoders: Pointer and Recursi
ve Transformer, and experiment with these models on a public dataset. We also pr
opose several metrics for measuring the accuracy of tree prediction and ground t
hese metrics in the domain of user experience. These contribute a new task and m
ethods to deep learning research.
****

Not All Features Are Equal: Feature Leveling Deep Neural Networks for Better Int
erpretation

Yingjing Lu,Runde Yang

Self-explaining models are models that reveal decision making parameters in an i
nterpretable manner so that the model reasoning process can be directly understo
od by human beings. General Linear Models (GLMs) are self-explaining because the
 model weights directly show how each feature contributes to the output value. H
owever, deep neural networks (DNNs) are in general not self-explaining due to th
e non-linearity of the activation functions, complex architectures, obscure feat
ure extraction and transformation process. In this work, we illustrate the fact
that existing deep architectures are hard to interpret because each hidden layer
 carries a mix of low level features and high level features. As a solution, we
propose a novel feature leveling architecture that isolates low level features f
rom high level features on a per-layer basis to better utilize the GLM layer in
the proposed architecture for interpretation. Experimental results show that our
 modified models are able to achieve competitive results comparing to main-strea
m architectures on standard datasets while being more self-explainable. Our impl
ementations and configurations are publicly available for reproductions.
****

Intrinsic Motivation for Encouraging Synergistic Behavior

Rohan Chitnis,Shubham Tulsiani,Saurabh Gupta,Abhinav Gupta

We study the role of intrinsic motivation as an exploration bias for reinforceme
nt learning in sparse-reward synergistic tasks, which are tasks where multiple a
gents must work together to achieve a goal they could not individually. Our key
idea is that a good guiding principle for intrinsic motivation in synergistic ta
sks is to take actions which affect the world in ways that would not be achieved
 if the agents were acting on their own. Thus, we propose to incentivize agents
to take (joint) actions whose effects cannot be predicted via a composition of t
he predicted effect for each individual agent. We study two instantiations of th
is idea, one based on the true states encountered, and another based on a dynami
cs model trained concurrently with the policy. While the former is simpler, the
latter has the benefit of being analytically differentiable with respect to the
action taken. We validate our approach in robotic bimanual manipulation and mult
i-agent locomotion tasks with sparse rewards; we find that our approach yields m
ore efficient learning than both 1) training with only the sparse reward and 2)
using the typical surprise-based formulation of intrinsic motivation, which does
 not bias toward synergistic behavior. Videos are available on the project webpa
ge: https://sites.google.com/view/iclr2020-synergistic.
****

Noisy Machines: Understanding noisy neural networks and enhancing robustness to
analog hardware errors using distillation

Chuteng Zhou,Prad Kadambi,Matthew Mattina,Paul N. Whatmough

The success of deep learning has brought forth a wave of interest in computer ha
rdware design to better meet the high demands of neural network inference. In pa
rticular, analog computing hardware has been heavily motivated specifically for
accelerating neural networks, based on either electronic, optical or photonic de
vices, which may well achieve lower power consumption than conventional digital
electronics. However, these proposed analog accelerators suffer from the intrins

ic noise generated by their physical components, which makes it challenging to a chieve high accuracy on deep neural networks. Hence, for successful deployment o n analog accelerators, it is essential to be able to train deep neural networks to be robust to random continuous noise in the network weights, which is a somew hat new challenge in machine learning. In this paper, we advance the understandi ng of noisy neural networks. We outline how a noisy neural network has reduced l earning capacity as a result of loss of mutual information between its input and output. To combat this, we propose using knowledge distillation combined with n oise injection during training to achieve more noise robust networks, which is d emonstrated experimentally across different networks and datasets, including Ima geNet. Our method achieves models with as much as 2X greater noise tolerance com pared with the previous best attempts, which is a significant step towards makin g analog hardware practical for deep learning.

****

Perceptual Regularization: Visualizing and Learning Generalizable Representation s

Hongzhou Lin,Joshua Robinson,Stefanie Jegelka

A deployable machine learning model relies on a good representation. Two desirab le criteria of a good representation are to be understandable, and to generalize to new tasks. We propose a technique termed perceptual regularization that enab les both visualization of the latent representation and control over the general ity of the learned representation. In particular our method provides a direct vi sualization of the effect that adversarial attacks have on the internal represen tation of a deep network. By visualizing the learned representation, we are also able to understand the attention of a model, obtaining visual evidence that sup ervised networks learn task-specific representations. We show models trained wit h perceptual regularization learn transferrable features, achieving significantl y higher accuracy in unseen tasks compared to standard supervised learning and m ulti-task methods.

****

Neural networks with motivation

Sergey A. Shuvaev,Ngoc B. Tran,Marcus Stephenson-Jones,Bo Li,Alexei A. Koulakov

How can animals behave effectively in conditions involving different motivationa l contexts? Here, we propose how reinforcement learning neural networks can lear n optimal behavior for dynamically changing motivational salience vectors. First , we show that Q-learning neural networks with motivation can navigate in enviro nment with dynamic rewards. Second, we show that such networks can learn complex behaviors simultaneously directed towards several goals distributed in an envir onment. Finally, we show that in Pavlovian conditioning task, the responses of t he neurons in our model resemble the firing patterns of neurons in the ventral p allidum (VP), a basal ganglia structure involved in motivated behaviors. We show that, similarly to real neurons, recurrent networks with motivation are compose d of two oppositely-tuned classes of neurons, responding to positive and negativ e rewards. Our model generates predictions for the VP connectivity. We conclude that networks with motivation can rapidly adapt their behavior to varying condit ions without changes in synaptic strength when expected reward is modulated by m otivation. Such networks may also provide a mechanism for how hierarchical reinf orcement learning is implemented in the brain.

****

RaCT: Toward Amortized Ranking-Critical Training For Collaborative Filtering

Sam Lobel*,Chunyuan Li*,Jianfeng Gao,Lawrence Carin

We investigate new methods for training collaborative filtering models based on actor-critic reinforcement learning, to more directly maximize ranking-based obj ective functions. Specifically, we train a critic network to approximate ranking -based metrics, and then update the actor network to directly optimize against t he learned metrics. In contrast to traditional learning-to-rank methods that req uire re-running the optimization procedure for new lists, our critic-based metho d amortizes the scoring process with a neural network, and can directly provide the (approximate) ranking scores for new lists.

We demonstrate the actor-critic's ability to significantly improve the performan
ce of a variety of prediction models, and achieve better or comparable performan
ce to a variety of strong baselines on three large-scale datasets.

****
ALBERT: A Lite BERT for Self-supervised Learning of Language Representations
Zhenzhong Lan,Mingda Chen,Sebastian Goodman,Kevin Gimpel,Piyush Sharma,Radu Sori
cut
Increasing model size when pretraining natural language representations often re
sults in improved performance on downstream tasks. However, at some point furthe
r model increases become harder due to GPU/TPU memory limitations and longer tra
ining times. To address these problems, we present two parameter-reduction tech
niques to lower memory consumption and increase the training speed of BERT~\cite
p{devlin2018bert}. Comprehensive empirical evidence shows that our proposed meth
ods lead to models that scale much better compared to the original BERT. We also
 use a self-supervised loss that focuses on modeling inter-sentence coherence, a
nd show it consistently helps downstream tasks with multi-sentence inputs. As a
result, our best model establishes new state-of-the-art results on the GLUE, RAC
E, and \squad benchmarks while having fewer parameters compared to BERT-large. T
he code and the pretrained models are available at https://github.com/google-res
earch/ALBERT.
****
Curriculum Learning for Deep Generative Models with Clustering
Deli Zhao,Jiapeng Zhu,Zhenfang Guo,Bo Zhang
Training generative models like Generative Adversarial Network (GAN) is challen
ging for noisy data. A novel curriculum learning algorithm pertaining to cluster
ing is proposed to address this issue in this paper. The curriculum construction
 is based on the centrality of underlying clusters in data points. The data poi
nts of high centrality takes priority of being fed into generative models during
 training. To make our algorithm scalable to large-scale data, the active set is
 devised, in the sense that every round of training proceeds only on an active s
ubset containing a small fraction of already trained data and the incremental da
ta of lower centrality. Moreover, the geometric analysis is presented to interpr
et the necessity of cluster curriculum for generative models. The experiments on
 cat and human-face data validate that our algorithm is able to learn the optima
l generative models (e.g. ProGAN) with respect to specified quality metrics for
noisy data. An interesting finding is that the optimal cluster curriculum is clo
sely related to the critical point of the geometric percolation process formulat
ed in the paper.
****
Sign-OPT: A Query-Efficient Hard-label Adversarial Attack
Minhao Cheng,Simranjit Singh,Patrick H. Chen,Pin-Yu Chen,Sijia Liu,Cho-Jui Hsieh
We study the most practical problem setup for evaluating adversarial robustness
of a machine learning system with limited access: the hard-label black-box atta
ck setting for generating adversarial examples, where limited model queries are
allowed and only the decision is provided to a queried data input. Several algor
ithms have been proposed for this problem but they typically require huge amount
 (>20,000) of queries for attacking one example. Among them, one of the state-of
-the-art approaches (Cheng et al., 2019) showed that hard-label attack can be mo
deled as an optimization problem where the objective function can be evaluated b
y binary search with additional model queries, thereby a zeroth order optimizati
on algorithm can be applied. In this paper, we adopt the same optimization formu
lation but propose to directly estimate the sign of gradient at any direction i
nstead of the gradient itself, which enjoys the benefit of single query.
Using this single query oracle for retrieving sign of directional derivative, we
 develop a novel query-efficient Sign-OPT approach for hard-label black-box atta
ck. We provide a convergence analysis of the new algorithm and conduct experimen
ts on several models on MNIST, CIFAR-10 and ImageNet.
We find that Sign-OPT attack consistently requires 5X to 10X fewer queries when
compared to the current state-of-the-art approaches, and usually converges to an

adversarial example with smaller perturbation.
****
Playing the lottery with rewards and multiple languages: lottery tickets in RL a
nd NLP
Haonan Yu,Sergey Edunov,Yuandong Tian,Ari S. Morcos
The lottery ticket hypothesis proposes that over-parameterization of deep neural
networks (DNNs) aids training by increasing the probability of a “lucky” sub-ne
twork initialization being present rather than by helping the optimization proce
ss (Frankle& Carbin, 2019). Intriguingly, this phenomenon suggests that initiali
zation strategies for DNNs can be improved substantially, but the lottery ticket
hypothesis has only previously been tested in the context of supervised learnin
g for natural image tasks. Here, we evaluate whether “winning ticket” initializa
tions exist in two different domains: natural language processing (NLP) and rein
forcement learning (RL).For NLP, we examined both recurrent LSTM models and larg
e-scale Transformer models (Vaswani et al., 2017). For RL, we analyzed a number
of discrete-action space tasks, including both classic control and pixel control
. Consistent with workin supervised image classification, we confirm that winnin
g ticket initializations generally outperform parameter-matched random initializ
ations, even at extreme pruning rates for both NLP and RL. Notably, we are able
to find winning ticket initializations for Transformers which enable models one-
third the size to achieve nearly equivalent performance. Together, these results
suggest that the lottery ticket hypothesis is not restricted to supervised lear
ning of natural images, but rather represents a broader phenomenon in DNNs.
****
Learning Space Partitions for Nearest Neighbor Search
Yihe Dong,Piotr Indyk,Ilya Razenshteyn,Tal Wagner
Space partitions of $\mathbb{R}^d$ underlie a vast and important
class of fast nearest neighbor search (NNS) algorithms. Inspired by recent theor
etical work on NNS for general metric spaces (Andoni et al. 2018b,c), we develop
a new framework for building space partitions reducing the problem to balanced
graph partitioning followed by supervised classification.
We instantiate this general approach with the KaHIP graph partitioner (Sanders a
nd Schulz 2013) and neural networks, respectively, to obtain a new partitioning
procedure called Neural Locality-Sensitive Hashing (Neural LSH). On several stan
dard benchmarks for NNS (Aumuller et al. 2017), our experiments show that the pa
rtitions obtained by Neural LSH consistently outperform partitions found by quan
tization-based and tree-based methods as well as classic, data-oblivious LSH.
****
Visual Interpretability Alone Helps Adversarial Robustness
Akhilan Boopathy,Sijia Liu,Gaoyuan Zhang,Pin-Yu Chen,Shiyu Chang,Luca Daniel
Recent works have empirically shown that there exist adversarial examples that
can be hidden from neural network interpretability, and interpretability is it
self susceptible to adversarial attacks. In this paper, we theoretically show t
hat with the correct measurement of interpretation, it is actually difficult to
hide adversarial examples, as confirmed by experiments on MNIST, CIFAR-10 and Re
stricted ImageNet. Spurred by that, we develop a novel defensive scheme built on
ly on robust interpretation (without resorting to adversarial loss minimization)
. We show that our defense achieves similar classification robustness to state-o
f-the-art robust training methods while attaining higher interpretation robustne
ss under various settings of adversarial attacks.
****
One-Shot Neural Architecture Search via Compressive Sensing
Minsu Cho,Mohammadreza Soltani,Chinmay Hegde
Neural architecture search (NAS), or automated design of neural network models,
remains a very challenging meta-learning problem. Several recent works (called "
one-shot" approaches) have focused on dramatically reducing NAS running time by
leveraging proxy models that still provide architectures with competitive perfor
mance. In our work, we propose a new meta-learning algorithm that we call CoNAS,
or Compressive sensing-based Neural Architecture Search. Our approach merges id
eas from one-shot NAS approaches with iterative techniques for learning low-degr

ee sparse Boolean polynomial functions. We validate our approach on several stan
dard test datasets, discover novel architectures hitherto unreported, and achiev
e competitive (or better) results in both performance and search time compared t
o existing NAS approaches. Further, we provide theoretical analysis via upper bo
unds on the number of validation error measurements needed to perform reliable m
eta-learning; to our knowledge, these analysis tools are novel to the NAS litera
ture and may be of independent interest.
****

End-to-end named entity recognition and relation extraction using pre-trained la
nguage models
John Giorgi,Xindi Wang,Nicola Sahar,Won Young Shin,Gary Bader,Bo Wang
Named entity recognition (NER) and relation extraction (RE) are two important ta
sks in information extraction and retrieval (IE & IR). Recent work has demonstra
ted that it is beneficial to learn these tasks jointly, which avoids the propaga
tion of error inherent in pipeline-based systems and improves performance. Howev
er, state-of-the-art joint models typically rely on external natural language pr
ocessing (NLP) tools, such as dependency parsers, limiting their usefulness to d
omains (e.g. news) where those tools perform well. The few neural, end-to-end mo
dels that have been proposed are trained almost completely from scratch. In this
 paper, we propose a neural, end-to-end model for jointly extracting entities an
d their relations which does not rely on external NLP tools and which integrates
 a large, pre-trained language model. Because the bulk of our model's parameters
 are pre-trained and we eschew recurrence for self-attention, our model is fast
to train. On 5 datasets across 3 domains, our model matches or exceeds state-of-
the-art performance, sometimes by a large margin.
****

How noise affects the Hessian spectrum in overparameterized neural networks
Mingwei Wei,David Schwab
Stochastic gradient descent (SGD) forms the core optimization method for deep ne
ural networks. While some theoretical progress has been made, it still remains u
nclear why SGD leads the learning dynamics in overparameterized networks to solu
tions that generalize well. Here we show that for overparameterized networks wit
h a degenerate valley in their loss landscape, SGD on average decreases the trac
e of the Hessian of the loss. We also generalize this result to other noise stru
ctures and show that isotropic noise in the non-degenerate subspace of the Hessi
an decreases its determinant. In addition to explaining SGDs role in sculpting t
he Hessian spectrum, this opens the door to new optimization approaches that may
 confer better generalization performance. We test our results with experiments
on toy models and deep neural networks.
****

A Simple Recurrent Unit with Reduced Tensor Product Representations
Shuai Tang,Paul Smolensky,Virginia R. de Sa
Widely used recurrent units, including Long-short Term Memory (LSTM) and Gated R
ecurrent Unit (GRU), perform well on natural language tasks, but their ability t
o learn structured representations is still questionable. Exploiting reduced Ten
sor Product Representations (TPRs) --- distributed representations of symbolic s
tructure in which vector-embedded symbols are bound to vector-embedded structura
l positions --- we propose the TPRU, a simple recurrent unit that, at each time
step, explicitly executes structural-role binding and unbinding operations to in
corporate structural information into learning. The gradient analysis of our pro
posed TPRU is conducted to support our model design, and its performance on mult
iple datasets shows the effectiveness of it. Furthermore, observations on lingui
stically grounded study demonstrate the interpretability of our TPRU.
****

Parallel Neural Text-to-Speech
Kainan Peng,Wei Ping,Zhao Song,Kexin Zhao
In this work, we first propose ParaNet, a non-autoregressive seq2seq model that
converts text to spectrogram. It is fully convolutional and obtains 46.7 times s
peed-up over Deep Voice 3 at synthesis while maintaining comparable speech quali
ty using a WaveNet vocoder. ParaNet also produces stable alignment between text

and speech on the challenging test sentences by iteratively improving the attent
ion in a layer-by-layer manner. Based on ParaNet, we build the first fully paral
lel neural text-to-speech system using parallel neural vocoders, which can synth
esize speech from text through a single feed-forward pass. We investigate sever
al parallel vocoders within the TTS system, including variants of IAF vocoders a
nd bipartite flow vocoder.
****

Context-Aware Object Detection With Convolutional Neural Networks
Yizhou Yan,Lei Cao,Samuel Madden,Elke Rundensteiner
Although the state-of-the-art object detection methods are successful in detecti
ng and classifying objects by leveraging deep convolutional neural networks (CNN
s), these methods overlook the semantic context which implies the probabilities
that different classes of objects occur jointly. In this work, we propose a cont
ext-aware CNN (or conCNN for short) that for the first time effectively enforces
 the semantics context constraints in the CNN-based object detector by leveragin
g the popular conditional random field (CRF) model in CNN. In particular, conCNN
 features a context-aware module that naturally models the mean-field inference
method for CRF using a stack of common CNN operations. It can be seamlessly plug
ged into any existing region-based object detection paradigm. Our experiments us
ing COCO datasets showcase that conCNN improves the average precision (AP) of ob
ject detection by 2 percentage points, while only introducing negligible extra t
raining overheads.
****

DeepV2D: Video to Depth with Differentiable Structure from Motion
Zachary Teed,Jia Deng
We propose DeepV2D, an end-to-end deep learning architecture for predicting dept
h from video. DeepV2D combines the representation ability of neural networks wi
th the geometric principles governing image formation. We compose a collection o
f classical geometric algorithms, which are converted into trainable modules and
 combined into an end-to-end differentiable architecture. DeepV2D interleaves tw
o stages: motion estimation and depth estimation. During inference, motion and d
epth estimation are alternated and converge to accurate depth.
****

TPO: TREE SEARCH POLICY OPTIMIZATION FOR CONTINUOUS ACTION SPACES
Amir Yazdanbakhsh,Ebrahim Songhori,Robert Ormandi,Anna Goldie,Azalia Mirhoseini
Monte Carlo Tree Search (MCTS) has achieved impressive results on a range of dis
crete environments, such as Go, Mario and Arcade games, but it has not yet fulfi
lled its true potential in continuous domains.In this work, we introduceTPO, a t
ree search based policy optimization method for continuous environments. TPO tak
es a hybrid approach to policy optimization. Building the MCTS tree in a contin
uous action space and updating the policy gradient using off-policy MCTS traject
ories are non-trivial. To overcome these challenges, we propose limiting tree se
arch branching factor by drawing only few action samples from the policy distrib
ution and define a new loss function based on the trajectories’ mean and standar
d deviations. Our approach led to some non-intuitive findings. MCTS training g
enerally requires a large number of samples and simulations. However, we observe
d that bootstrappingtree search with a pre-trained policy allows us to achieve h
igh quality results with a low MCTS branching factor and few number of simulatio
ns. Without the proposed policy bootstrapping, continuous MCTS would require a m
uch larger branching factor and simulation count, rendering it computationally a
nd prohibitively expensive. In our experiments, we use PPO as our baseline polic
y optimization algorithm. TPO significantly improves the policy on nearly all of
 our benchmarks. For example, in complex environments such as Humanoid, we achi
eve a 2.5×improvement over the baseline algorithm.
****

Gaussian Process Meta-Representations Of Neural Networks
Theofanis Karaletsos,Thang Bui
Bayesian inference offers a theoretically grounded and general way to train neur
al networks and can potentially give calibrated uncertainty. It is, however, cha
llenging to specify a meaningful and tractable prior over the network parameters

. More crucially, many existing inference methods assume mean-field approximate posteriors, ignoring interactions between parameters in high-dimensional weight space. To this end, this paper introduces two innovations: (i) a Gaussian process-based hierarchical model for the network parameters based on recently introduced unit embeddings that can flexibly encode weight structures, and (ii) input-dependent contextual variables for the weight prior that can provide convenient ways to regularize the function space being modeled by the NN through the use of kernels.
Furthermore, we develop an efficient structured variational inference scheme that alleviates the need to perform inference in the weight space whilst retaining and learning non-trivial correlations between network parameters.
We show these models provide desirable test-time uncertainty estimates, demonstrate cases of modeling inductive biases for neural networks with kernels and demonstrate competitive predictive performance of the proposed model and algorithm over alternative approaches on a range of classification and active learning tasks.
****

CAN ALTQ LEARN FASTER: EXPERIMENTS AND THEORY
Bowen Weng,Huaqing Xiong,Yingbin Liang,Wei Zhang
Differently from the popular Deep Q-Network (DQN) learning, Alternating Q-learning (AltQ) does not fully fit a target Q-function at each iteration, and is generally known to be unstable and inefficient. Limited applications of AltQ mostly rely on substantially altering the algorithm architecture in order to improve its performance. Although Adam appears to be a natural solution, its performance in AltQ has rarely been studied before. In this paper, we first provide a solid exploration on how well AltQ performs with Adam. We then take a further step to improve the implementation by adopting the technique of parameter restart. More specifically, the proposed algorithms are tested on a batch of Atari 2600 games and exhibit superior performance than the DQN learning method. The convergence rate of the slightly modified version of the proposed algorithms is characterized under the linear function approximation. To the best of our knowledge, this is the first theoretical study on the Adam-type algorithms in Q-learning.
****

The Break-Even Point on Optimization Trajectories of Deep Neural Networks
Stanislaw Jastrzebski,Maciej Szymczak,Stanislav Fort,Devansh Arpit,Jacek Tabor,Kyunghyun Cho*,Krzysztof Geras*
The early phase of training of deep neural networks is critical for their final performance. In this work, we study how the hyperparameters of stochastic gradient descent (SGD) used in the early phase of training affect the rest of the optimization trajectory. We argue for the existence of the "``break-even" point on this trajectory, beyond which the curvature of the loss surface and noise in the gradient are implicitly regularized by SGD. In particular, we demonstrate on multiple classification tasks that using a large learning rate in the initial phase of training reduces the variance of the gradient, and improves the conditioning of the covariance of gradients. These effects are beneficial from the optimization perspective and become visible after the break-even point. Complementing prior work, we also show that using a low learning rate results in bad conditioning of the loss surface even for a neural network with batch normalization layers. In short, our work shows that key properties of the loss surface are strongly influenced by SGD in the early phase of training. We argue that studying the impact of the identified effects on generalization is a promising future direction.
****

Towards Better Understanding of Adaptive Gradient Algorithms in Generative Adversarial Nets
Mingrui Liu,Youssef Mroueh,Jerret Ross,Wei Zhang,Xiaodong Cui,Payel Das,Tianbao Yang
Adaptive gradient algorithms perform gradient-based updates using the history of gradients and are ubiquitous in training deep neural networks. While adaptive gradient methods theory is well understood for minimization problems, the underlying factors driving their empirical success in min-max problems such as GANs rem

ain unclear. In this paper, we aim at bridging this gap from both theoretical a
nd empirical perspectives. First, we analyze a variant of Optimistic Stochastic
Gradient (OSG) proposed in~\citep{daskalakis2017training} for solving a class of
 non-convex non-concave min-max problem and establish $O(\epsilon^{-4})$ complex
ity for finding ϵ-first-order stationary point, in which the algorithm
only requires invoking one stochastic first-order oracle while enjoying state-of
-the-art iteration complexity achieved by stochastic extragradient method by~\ci
tep{iusem2017extragradient}. Then we propose an adaptive variant of OSG named Op
timistic Adagrad (OAdagrad) and reveal an \emph{improved} adaptive complexity $O
\left(\epsilon^{-\frac{2}{1-\alpha}}\right)$, where α characterizes the g
rowth rate of the cumulative stochastic gradient and $0\leq \alpha\leq 1/2$. To
the best of our knowledge, this is the first work for establishing adaptive comp
lexity in non-convex non-concave min-max optimization. Empirically, our experime
nts show that indeed adaptive gradient algorithms outperform their non-adaptive
counterparts in GAN training. Moreover, this observation can be explained by the
 slow growth rate of the cumulative stochastic gradient, as observed empirically
.
****

Exploration Based Language Learning for Text-Based Games
Andrea Madotto,Mahdi Namazifar,Joost Huizinga,Piero Molino,Adrien Ecoffet,Huaixi
u Zheng,Alexandros Papangelis,Dian Yu,Chandra Khatri,Gokhan Tur
This work presents an exploration and imitation-learning-based agent capable of
state-of-the-art performance in playing text-based computer games. Text-based co
mputer games describe their world to the player through natural language and exp
ect the player to interact with the game using text. These games are of interest
 as they can be seen as a testbed for language understanding, problem-solving, a
nd language generation by artificial agents. Moreover, they provide a learning e
nvironment in which these skills can be acquired through interactions with an en
vironment rather than using fixed corpora.
One aspect that makes these games particularly challenging for learning agents i
s the combinatorially large action space.
Existing methods for solving text-based games are limited to games that are eith
er very simple or have an action space restricted to a predetermined set of admi
ssible actions. In this work, we propose to use the exploration approach of Go-E
xplore (Ecoffet et al., 2019) for solving text-based games. More specifically, i
n an initial exploration phase, we first extract trajectories with high rewards,
 after which we train a policy to solve the game by imitating these trajectories
.
Our experiments show that this approach outperforms existing solutions in solvin
g text-based games, and it is more sample efficient in terms of the number of in
teractions with the environment. Moreover, we show that the learned policy can g
eneralize better than existing solutions to unseen games without using any restr
iction on the action space.
****

Robust And Interpretable Blind Image Denoising Via Bias-Free Convolutional Neura
l Networks
Sreyas Mohan,Zahra Kadkhodaie,Eero P. Simoncelli,Carlos Fernandez-Granda
We study the generalization properties of deep convolutional neural networks for
 image denoising in the presence of varying noise levels. We provide extensive e
mpirical evidence that current state-of-the-art architectures systematically ove
rfit to the noise levels in the training set, performing very poorly at new nois
e levels. We show that strong generalization can be achieved through a simple ar
chitectural modification: removing all additive constants. The resulting "bias-f
ree" networks attain state-of-the-art performance over a broad range of noise le
vels, even when trained over a limited range. They are also locally linear, whic
h enables direct analysis with linear-algebraic tools. We show that the denoisi
ng map can be visualized locally as a filter that adapts to both image structure
 and noise level. In addition, our analysis reveals that deep networks implicitl
y perform a projection onto an adaptively-selected low-dimensional subspace, wit
h dimensionality inversely proportional to noise level, that captures features o

f natural images.
****
CM3: Cooperative Multi-goal Multi-stage Multi-agent Reinforcement Learning
Jiachen Yang,Alireza Nakhaei,David Isele,Kikuo Fujimura,Hongyuan Zha
A variety of cooperative multi-agent control problems require agents to achieve individual goals while contributing to collective success. This multi-goal multi-agent setting poses difficulties for recent algorithms, which primarily target settings with a single global reward, due to two new challenges: efficient exploration for learning both individual goal attainment and cooperation for others' success, and credit-assignment for interactions between actions and goals of different agents. To address both challenges, we restructure the problem into a novel two-stage curriculum, in which single-agent goal attainment is learned prior to learning multi-agent cooperation, and we derive a new multi-goal multi-agent policy gradient with a credit function for localized credit assignment. We use a function augmentation scheme to bridge value and policy functions across the curriculum. The complete architecture, called CM3, learns significantly faster than direct adaptations of existing algorithms on three challenging multi-goal multi-agent problems: cooperative navigation in difficult formations, negotiating multi-vehicle lane changes in the SUMO traffic simulator, and strategic cooperation in a Checkers environment.
****
Deep Imitative Models for Flexible Inference, Planning, and Control
Nicholas Rhinehart,Rowan McAllister,Sergey Levine
Imitation Learning (IL) is an appealing approach to learn desirable autonomous behavior. However, directing IL to achieve arbitrary goals is difficult. In contrast, planning-based algorithms use dynamics models and reward functions to achieve goals. Yet, reward functions that evoke desirable behavior are often difficult to specify. In this paper, we propose "Imitative Models" to combine the benefits of IL and goal-directed planning. Imitative Models are probabilistic predictive models of desirable behavior able to plan interpretable expert-like trajectories to achieve specified goals. We derive families of flexible goal objectives, including constrained goal regions, unconstrained goal sets, and energy-based goals. We show that our method can use these objectives to successfully direct behavior. Our method substantially outperforms six IL approaches and a planning-based approach in a dynamic simulated autonomous driving task, and is efficiently learned from expert demonstrations without online data collection. We also show our approach is robust to poorly-specified goals, such as goals on the wrong side of the road.
****
Exploiting Excessive Invariance caused by Norm-Bounded Adversarial Robustness
Jörn-Henrik Jacobsen,Jens Behrmann,Nicholas Carlini,Florian Tramèr,Nicolas Papernot
Adversarial examples are malicious inputs crafted to cause a model to misclassify them. In their most common instantiation, "perturbation-based" adversarial examples introduce changes to the input that leave its true label unchanged, yet result in a different model prediction. Conversely, "invariance-based" adversarial examples insert changes to the input that leave the model's prediction unaffected despite the underlying input's label having changed. So far, the relationship between these two notions of adversarial examples has not been studied, we close this gap.

We demonstrate that solely achieving perturbation-based robustness is insufficient for complete adversarial robustness. Worse, we find that classifiers trained to be Lp-norm robust are more vulnerable to invariance-based adversarial examples than their undefended counterparts. We construct theoretical arguments and analytical examples to justify why this is the case. We then illustrate empirically that the consequences of excessive perturbation-robustness can be exploited to craft new attacks. Finally, we show how to attack a provably robust defense --- certified on the MNIST test set to have at least 87% accuracy (with respect to the original test labels) under perturbations of Linfinity-norm below epsilon=0.4

--- and reduce its accuracy (under this threat model with respect to an ensembl
e of human labelers) to 60% with an automated attack, or just 12% with human-cra
fted adversarial examples.
****

Defective Convolutional Layers Learn Robust CNNs
Tiange Luo,Tianle Cai,Xiaomeng Zhang,Siyu Chen,Di He,Liwei Wang
Robustness of convolutional neural networks has recently been highlighted by the
 adversarial examples, i.e., inputs added with well-designed perturbations which
 are imperceptible to humans but can cause the network to give incorrect outputs
. Recent research suggests that the noises in adversarial examples break the tex
tural structure, which eventually leads to wrong predictions by convolutional ne
ural networks. To help a convolutional neural network make predictions relying l
ess on textural information, we propose defective convolutional layers which con
tain defective neurons whose activations are set to be a constant function. As t
he defective neurons contain no information and are far different from the stand
ard neurons in its spatial neighborhood, the textural features cannot be accurat
ely extracted and the model has to seek for other features for classification, s
uch as the shape. We first show that predictions made by the defective CNN are l
ess dependent on textural information, but more on shape information, and furthe
r find that adversarial examples generated by the defective CNN appear to have s
emantic shapes. Experimental results demonstrate the defective CNN has higher de
fense ability than the standard CNN against various types of attack. In particul
ar, it achieves state-of-the-art performance against transfer-based attacks with
out applying any adversarial training.
****

DASGrad: Double Adaptive Stochastic Gradient
Kin Gutierrez,Cristian Challu,Jin Li,Artur Dubrawski
Adaptive moment methods have been remarkably successful for optimization under t
he presence of high dimensional or sparse gradients, in parallel to this, adapti
ve sampling probabilities for SGD have allowed optimizers to improve convergence
 rates by prioritizing examples to learn efficiently. Numerous applications in t
he past have implicitly combined adaptive moment methods with adaptive probabili
ties yet the theoretical guarantees of such procedures have not been explored. W
e formalize double adaptive stochastic gradient methods DASGrad as an optimizati
on technique and analyze its convergence improvements in a stochastic convex opt
imization setting, we provide empirical validation of our findings with convex a
nd non convex objectives. We observe that the benefits of the method increase wi
th the model complexity and variability of the gradients, and we explore the res
ulting utility in extensions to transfer learning.
****

Finding Mixed Strategy Nash Equilibrium for Continuous Games through Deep Learni
ng
Zehao Dou,Xiang Yan,Dongge Wang,Xiaotie Deng
Nash equilibrium has long been a desired solution concept in multi-player games,
 especially for those on continuous strategy spaces, which have attracted a rapi
dly growing amount of interests due to advances in research applications such as
 the generative adversarial networks. Despite the fact that several deep learnin
g based approaches are designed to obtain pure strategy Nash equilibrium, it is
rather luxurious to assume the existence of such an equilibrium. In this paper,
we present a new method to approximate mixed strategy Nash equilibria in multi-p
layer continuous games, which always exist and include the pure ones as a specia
l case. We remedy the pure strategy weakness by adopting the pushforward measure
 technique to represent a mixed strategy in continuous spaces. That allows us to
 generalize the Gradient-based Nikaido-Isoda (GNI) function to measure the dista
nce between the players' joint strategy profile and a Nash equilibrium. Applying
 the gradient descent algorithm, our approach is shown to converge to a stationa
ry Nash equilibrium under the convexity assumption on payoff functions, the same
 popular setting as in previous studies.
In numerical experiments, our method consistently and significantly outperforms
 recent works on approximating Nash equilibrium for quadratic games, general blo

tto games, and GAMUT games.
****
The Logical Expressiveness of Graph Neural Networks
Pablo Barceló,Egor V. Kostylev,Mikael Monet,Jorge Pérez,Juan Reutter,Juan Pablo Silva
The ability of graph neural networks (GNNs) for distinguishing nodes in graphs has been recently characterized in terms of the Weisfeiler-Lehman (WL) test for checking graph isomorphism. This characterization, however, does not settle the issue of which Boolean node classifiers (i.e., functions classifying nodes in graphs as true or false) can be expressed by GNNs. We tackle this problem by focusing on Boolean classifiers expressible as formulas in the logic FOC2, a well-studied fragment of first order logic. FOC2 is tightly related to the WL test, and hence to GNNs. We start by studying a popular class of GNNs, which we call AC-GNNs, in which the features of each node in the graph are updated, in successive layers, only in terms of the features of its neighbors. We show that this class of GNNs is too weak to capture all FOC2 classifiers, and provide a syntactic characterization of the largest subclass of FOC2 classifiers that can be captured by AC-GNNs. This subclass coincides with a logic heavily used by the knowledge representation community. We then look at what needs to be added to AC-GNNs for capturing all FOC2 classifiers. We show that it suffices to add readout functions, which allow to update the features of a node not only in terms of its neighbors, but also in terms of a global attribute vector. We call GNNs of this kind ACR-GNNs. We experimentally validate our findings showing that, on synthetic data conforming to FOC2 formulas, AC-GNNs struggle to fit the training data while ACR-GNNs can generalize even to graphs of sizes not seen during training.
****
Pre-training Tasks for Embedding-based Large-scale Retrieval
Wei-Cheng Chang,Felix X. Yu,Yin-Wen Chang,Yiming Yang,Sanjiv Kumar
We consider the large-scale query-document retrieval problem: given a query (e.g., a question), return the set of relevant documents (e.g., paragraphs containing the answer) from a large document corpus. This problem is often solved in two steps. The retrieval phase first reduces the solution space, returning a subset of candidate documents. The scoring phase then re-ranks the documents. Critically, the retrieval algorithm not only desires high recall but also requires to be highly efficient, returning candidates in time sublinear to the number of documents. Unlike the scoring phase witnessing significant advances recently due to the BERT-style pre-training tasks on cross-attention models, the retrieval phase remains less well studied. Most previous works rely on classic Information Retrieval (IR) methods such as BM-25 (token matching + TF-IDF weights). These models only accept sparse handcrafted features and can not be optimized for different downstream tasks of interest. In this paper, we conduct a comprehensive study on the embedding-based retrieval models. We show that the key ingredient of learning a strong embedding-based Transformer model is the set of pre-training tasks. With adequately designed paragraph-level pre-training tasks, the Transformer models can remarkably improve over the widely-used BM-25 as well as embedding models without Transformers. The paragraph-level pre-training tasks we studied are Inverse Cloze Task (ICT), Body First Selection (BFS), Wiki Link Prediction (WLP), and the combination of all three.
****
The Benefits of Over-parameterization at Initialization in Deep ReLU Networks
Devansh Arpit,Yoshua Bengio
It has been noted in existing literature that over-parameterization in ReLU networks generally improves performance. While there could be several factors involved behind this, we prove some desirable theoretical properties at initialization which may be enjoyed by ReLU networks. Specifically, it is known that He initialization in deep ReLU networks asymptotically preserves variance of activations in the forward pass and variance of gradients in the backward pass for infinitely wide networks, thus preserving the flow of information in both directions. Our paper goes beyond these results and shows novel properties that hold under He initialization: i) the norm of hidden activation of each layer is equal to the no

rm of the input, and, ii) the norm of weight gradient of each layer is equal to the product of norm of the input vector and the error at output layer. These results are derived using the PAC analysis framework, and hold true for finitely sized datasets such that the width of the ReLU network only needs to be larger than a certain finite lower bound. As we show, this lower bound depends on the depth of the network and the number of samples, and by the virtue of being a lower bound, over-parameterized ReLU networks are endowed with these desirable properties. For the aforementioned hidden activation norm property under He initialization, we further extend our theory and show that this property holds for a finite width network even when the number of data samples is infinite. Thus we overcome several limitations of existing papers, and show new properties of deep ReLU networks at initialization.
****

A Training Scheme for the Uncertain Neuromorphic Computing Chips
Qingtian Zhang,Bin Gao,Huaqiang Wu
Uncertainty is a very important feature of the intelligence and helps the brain become a flexible, creative and powerful intelligent system. The crossbar-based neuromorphic computing chips, in which the computing is mainly performed by analog circuits, have the uncertainty and can be used to imitate the brain. However, most of the current deep neural networks have not taken the uncertainty of the neuromorphic computing chip into consideration. Therefore, their performances on the neuromorphic computing chips are not as good as on the original platforms (CPUs/GPUs). In this work, we proposed the uncertainty adaptation training scheme (UATS) that tells the uncertainty to the neural network in the training process. The experimental results show that the neural networks can achieve comparable inference performances on the uncertain neuromorphic computing chip compared to the results on the original platforms, and much better than the performances without this training scheme.
****

Mildly Overparametrized Neural Nets can Memorize Training Data Efficiently
Rong Ge,Runzhe Wang,Haoyu Zhao
It has been observed \citep{zhang2016understanding} that deep neural networks can memorize: they achieve 100\% accuracy on training data. Recent theoretical results explained such behavior in highly overparametrized regimes, where the number of neurons in each layer is larger than the number of training samples. In this paper, we show that neural networks can be trained to memorize training data perfectly in a mildly overparametrized regime, where the number of parameters is just a constant factor more than the number of training samples, and the number of neurons is much smaller.
****

Deep Graph Translation
Xiaojie Guo,Lingfei Wu,Liang Zhao
Deep graph generation models have achieved great successes recently, among which, however, are typically unconditioned generative models that have no control over the target graphs are given an input graph. In this paper, we propose a novel Graph-Translation-Generative-Adversarial-Networks (GT-GAN) that transforms the input graphs into their target output graphs. GT-GAN consists of a graph translator equipped with innovative graph convolution and deconvolution layers to learn the translation mapping considering both global and local features, and a new conditional graph discriminator to classify target graphs by conditioning on input graphs. Extensive experiments on multiple synthetic and real-world datasets demonstrate that our proposed GT-GAN significantly outperforms other baseline methods in terms of both effectiveness and scalability. For instance, GT-GAN achieves at least 10X and 15X faster runtimes than GraphRNN and RandomVAE, respectively, when the size of the graph is around 50.
****

Are Transformers universal approximators of sequence-to-sequence functions?
Chulhee Yun,Srinadh Bhojanapalli,Ankit Singh Rawat,Sashank Reddi,Sanjiv Kumar
Despite the widespread adoption of Transformer models for NLP tasks, the expressive power of these models is not well-understood. In this paper, we establish th

at Transformer models are universal approximators of continuous permutation equivariant sequence-to-sequence functions with compact support, which is quite surprising given the amount of shared parameters in these models. Furthermore, using positional encodings, we circumvent the restriction of permutation equivariance, and show that Transformer models can universally approximate arbitrary continuous sequence-to-sequence functions on a compact domain. Interestingly, our proof techniques clearly highlight the different roles of the self-attention and the feed-forward layers in Transformers. In particular, we prove that fixed width self-attention layers can compute contextual mappings of the input sequences, playing a key role in the universal approximation property of Transformers. Based on this insight from our analysis, we consider other simpler alternatives to self-attention layers and empirically evaluate them.
****

Meta-Dataset: A Dataset of Datasets for Learning to Learn from Few Examples
Eleni Triantafillou,Tyler Zhu,Vincent Dumoulin,Pascal Lamblin,Utku Evci,Kelvin Xu,Ross Goroshin,Carles Gelada,Kevin Swersky,Pierre-Antoine Manzagol,Hugo Larochelle
Few-shot classification refers to learning a classifier for new classes given only a few examples. While a plethora of models have emerged to tackle it, we find the procedure and datasets that are used to assess their progress lacking. To address this limitation, we propose Meta-Dataset: a new benchmark for training and evaluating models that is large-scale, consists of diverse datasets, and presents more realistic tasks. We experiment with popular baselines and meta-learners on Meta-Dataset, along with a competitive method that we propose. We analyze performance as a function of various characteristics of test tasks and examine the models' ability to leverage diverse training sources for improving their generalization. We also propose a new set of baselines for quantifying the benefit of meta-learning in Meta-Dataset. Our extensive experimentation has uncovered important research challenges and we hope to inspire work in these directions.
****

Decoupling Weight Regularization from Batch Size for Model Compression
Dongsoo Lee,Se Jung Kwon,Byeongwook Kim,Yongkweon Jeon,Baeseong Park,Jeongin Yun,Gu-Yeon Wei
Conventionally, compression-aware training performs weight compression for every mini-batch to compute the impact of compression on the loss function. In this paper, in order to study when would be the right time to compress weights during optimization steps, we propose a new hyper-parameter called Non-Regularization period or NR period during which weights are not updated for regularization. We first investigate the influence of NR period on regularization using weight decay and weight random noise insertion. Throughout various experiments, we show that stronger weight regularization demands longer NR period (regardless of batch size) to best utilize regularization effects. From our empirical evidence, we argue that weight regularization for every mini-batch allows small weight updates only and limited regularization effects such that there is a need to search for right NR period and weight regularization strength to enhance model accuracy. Consequently, NR period becomes especially crucial for model compression where large weight updates are necessary to increase compression ratio. Using various models, we show that simple weight updates to comply with compression formats along with long NR period is enough to achieve high compression ratio and model accuracy.
****

Zero-Shot Out-of-Distribution Detection with Feature Correlations
Chandramouli S Sastry,Sageev Oore
When presented with Out-of-Distribution (OOD) examples, deep neural networks yield confident, incorrect predictions. Detecting OOD examples is challenging, and the potential risks are high. In this paper, we propose to detect OOD examples by identifying inconsistencies between activity patterns and class predicted. We find that characterizing activity patterns by feature correlations and identifying anomalies in pairwise feature correlation values can yield high OOD detection rates. We identify anomalies in the pairwise feature correlations by simply com

paring each pairwise correlation value with its respective range observed over t
he training data. Unlike many approaches, this can be used with any pre-trained
softmax classifier and does not require access to OOD data for fine-tuning hyper
parameters, nor does it require OOD access for inferring parameters. The method
is applicable across a variety of architectures and vision datasets and generall
y performs better than or equal to state-of-the-art OOD detection methods, inclu
ding those that do assume access to OOD examples.
****

Proactive Sequence Generator via Knowledge Acquisition
Qing Sun,James Cross,Dmitriy Genzel
Sequence-to-sequence models such as transformers, which are now being used in a
wide variety of NLP tasks, typically need to have very high capacity in order to
 perform well. Unfortunately, in production, memory size and inference speed are
 all strictly constrained. To address this problem, Knowledge Distillation (KD)
, a technique to train small models to mimic larger pre-trained models, has draw
n lots of attention. The KD approach basically attempts to maximize recall, i.e
., ranking Top-k”tokens in teacher models as higher as possible, however, wherea
s precision is more important for sequence generation because of exposure bias.
 Motivated by this, we develop Knowledge Acquisition (KA) where student models r
eceive log q(y_t|y_{<t},x) as rewards when producing the next token y_t given pr
evious tokens y_{<t} and the source sentence x. We demonstrate the effectivene
ss of our approach on WMT’17 De-En and IWSLT’15 Th-En translation tasks, with ex
perimental results showing that our approach gains +0.7-1.1 BLEU score compared
to token-level knowledge distillation.
****

Multi-scale Attributed Node Embedding
Benedek Rozemberczki,Carl Allen,Rik Sarkar
We present network embedding algorithms that capture information about a node fr
om the local distribution over node attributes around it, as observed over rando
m walks following an approach similar to Skip-gram. Observations from neighborho
ods of different sizes are either pooled (AE) or encoded distinctly in a multi-s
cale approach (MUSAE). Capturing attribute-neighborhood relationships over mult
iple scales is useful for a diverse range of applications, including latent feat
ure identification across disconnected networks with similar attributes. We prov
e theoretically that matrices of node-feature pointwise mutual information are i
mplicitly factorized by the embeddings. Experiments show that our algorithms are
 robust, computationally efficient and outperform comparable models on social, w
eb and citation network datasets.
****

Understanding the functional and structural differences across excitatory and in
hibitory neurons
Sun Minni,Li Ji-An,Theodore Moskovitz,Grace Lindsay,Kenneth Miller,Mario Dipoppa
,Guangyu Robert Yang
One of the most fundamental organizational principles of the brain is the separa
tion of excitatory (E) and inhibitory (I) neurons. In addition to their opposing
 effects on post-synaptic neurons, E and I cells tend to differ in their selecti
vity and connectivity. Although many such differences have been characterized ex
perimentally, it is not clear why they exist in the first place. We studied this
 question in deep networks equipped with E and I cells. We found that salient di
stinctions between E and I neurons emerge across various deep convolutional recu
rrent networks trained to perform standard object classification tasks. We explo
red the necessary conditions for the networks to develop distinct selectivity an
d connectivity across cell types. We found that neurons that project to higher-o
rder areas will have greater stimulus selectivity, regardless of whether they ar
e excitatory or not. Sparser connectivity is required for higher selectivity, bu
t only when the recurrent connections are excitatory. These findings demonstrate
 that the functional and structural differences observed across E and I neurons
are not independent, and can be explained using a smaller number of factors.
****

One-Shot Pruning of Recurrent Neural Networks by Jacobian Spectrum Evaluation

Shunshi Zhang,Bradly C. Stadie

Recent advances in the sparse neural network literature have made it possible to prune many large feed forward and convolutional networks with only a small quantity of data. Yet, these same techniques often falter when applied to the problem of recovering sparse recurrent networks. These failures are quantitative: when pruned with recent techniques, RNNs typically obtain worse performance than they do under a simple random pruning scheme. The failures are also qualitative: the distribution of active weights in a pruned LSTM or GRU network tend to be concentrated in specific neurons and gates, and not well dispersed across the entire architecture. We seek to rectify both the quantitative and qualitative issues with recurrent network pruning by introducing a new recurrent pruning objective derived from the spectrum of the recurrent Jacobian. Our objective is data efficient (requiring only 64 data points to prune the network), easy to implement, and produces 95 % sparse GRUs that significantly improve on existing baselines. We evaluate on sequential MNIST, Billion Words, and Wikitext.

****

Differentially Private Meta-Learning

Jeffrey Li,Mikhail Khodak,Sebastian Caldas,Ameet Talwalkar

Parameter-transfer is a well-known and versatile approach for meta-learning, with applications including few-shot learning, federated learning, with personalization, and reinforcement learning. However, parameter-transfer algorithms often require sharing models that have been trained on the samples from specific tasks, thus leaving the task-owners susceptible to breaches of privacy. We conduct the first formal study of privacy in this setting and formalize the notion of task-global differential privacy as a practical relaxation of more commonly studied threat models. We then propose a new differentially private algorithm for gradient-based parameter transfer that not only satisfies this privacy requirement but also retains provable transfer learning guarantees in convex settings. Empirically, we apply our analysis to the problems of federated learning with personalization and few-shot classification, showing that allowing the relaxation to task-global privacy from the more commonly studied notion of local privacy leads to dramatically increased performance in recurrent neural language modeling and image classification.

****

Leveraging Adversarial Examples to Obtain Robust Second-Order Representations

Mohit Prabhushankar,Gukyeong Kwon,Dogancan Temel,Ghassan AlRegib

Deep neural networks represent data as projections on trained weights in a high dimensional manifold. This is a first-order based absolute representation that is widely used due to its interpretable nature and simple mathematical functionality. However, in the application of visual recognition, first-order representations trained on pristine images have shown a vulnerability to distortions. Visual distortions including imaging acquisition errors and challenging environmental conditions like blur, exposure, snow and frost cause incorrect classification in first-order neural nets. To eliminate vulnerabilities under such distortions, we propose representing data points by their relative positioning in a high dimensional manifold instead of their absolute positions. Such a positioning scheme is based on a data point’s second-order property. We obtain a data point’s second-order representation by creating adversarial examples to all possible decision boundaries and tracking the movement of corresponding boundaries. We compare our representation against first-order methods and show that there is an increase of more than 14% under severe distortions for ResNet-18. We test the generalizability of the proposed representation on larger networks and on 19 complex and real-world distortions from CIFAR-10-C. Furthermore, we show how our proposed representation can be used as a plug-in approach on top of any network. We also provide methodologies to scale our proposed representation to larger datasets.

****

CLEVRER: Collision Events for Video Representation and Reasoning

Kexin Yi*,Chuang Gan*,Yunzhu Li,Pushmeet Kohli,Jiajun Wu,Antonio Torralba,Joshua B. Tenenbaum

The ability to reason about temporal and causal events from videos lies at the c

ore of human intelligence. Most video reasoning benchmarks, however, focus on pa ttern recognition from complex visual and language input, instead of on causal s tructure. We study the complementary problem, exploring the temporal and causal structures behind videos of objects with simple visual appearance. To this end, we introduce the CoLlision Events for Video REpresentation and Reasoning (CLEVRE R) dataset, a diagnostic video dataset for systematic evaluation of computation al models on a wide range of reasoning tasks. Motivated by the theory of human casual judgment, CLEVRER includes four types of question: descriptive (e.g., 'w hat color'), explanatory ('what's responsible for'), predictive ('what will happ en next'), and counterfactual ('what if'). We evaluate various state-of-the-art models for visual reasoning on our benchmark. While these models thrive on the perception-based task (descriptive), they perform poorly on the causal tasks (ex planatory, predictive and counterfactual), suggesting that a principled approach for causal reasoning should incorporate the capability of both perceiving compl ex visual and language inputs, and understanding the underlying dynamics and cau sal relations. We also study an oracle model that explicitly combines these comp onents via symbolic representations.
****

Using Logical Specifications of Objectives in Multi-Objective Reinforcement Lear ning

Kolby Nottingham,Anand Balakrishnan,Jyotirmoy Deshmukh,Connor Christopherson,Dav id Wingate

In the multi-objective reinforcement learning (MORL) paradigm, the relative impo rtance of each environment objective is often unknown prior to training, so agen ts must learn to specialize their behavior to optimize different combinations of environment objectives that are specified post-training. These are typically li near combinations, so the agent is effectively parameterized by a weight vector that describes how to balance competing environment objectives. However, many re al world behaviors require non-linear combinations of objectives. Additionally, the conversion between desired behavior and weightings is often unclear.
In this work, we explore the use of a language based on propositional logic with quantitative semantics--in place of weight vectors--for specifying non-linear b ehaviors in an interpretable way. We use a recurrent encoder to encode logical c ombinations of objectives, and train a MORL agent to generalize over these encod ings. We test our agent in several grid worlds with various objectives and show that our agent can generalize to many never-before-seen specifications with perf ormance comparable to single policy baseline agents. We also demonstrate our age nt's ability to generate meaningful policies when presented with novel specifica tions and quickly specialize to novel specifications.
****

Efficient Training of Robust and Verifiable Neural Networks

Akhilan Boopathy,Lily Weng,Sijia Liu,Pin-Yu Chen,Luca Daniel

Recent works have developed several methods of defending neural networks against adversarial attacks with certified guarantees. We propose that many common cert ified defenses can be viewed under a unified framework of regularization. This u nified framework provides a technique for comparing different certified defenses with respect to robust generalization. In addition, we develop a new regularize r that is both more efficient than existing certified defenses and can be used t o train networks with higher certified accuracy. Our regularizer also extends to an L0 threat model and ensemble models. Through experiments on MNIST, CIFAR-10 and GTSRB, we demonstrate improvements in training speed and certified accuracy compared to state-of-the-art certified defenses.
****

Learning Compositional Koopman Operators for Model-Based Control

Yunzhu Li,Hao He,Jiajun Wu,Dina Katabi,Antonio Torralba

Finding an embedding space for a linear approximation of a nonlinear dynamical s ystem enables efficient system identification and control synthesis. The Koopman operator theory lays the foundation for identifying the nonlinear-to-linear coo rdinate transformations with data-driven methods. Recently, researchers have pro posed to use deep neural networks as a more expressive class of basis functions

for calculating the Koopman operators. These approaches, however, assume a fixed dimensional state space; they are therefore not applicable to scenarios with a variable number of objects. In this paper, we propose to learn compositional Koopman operators, using graph neural networks to encode the state into object-centric embeddings and using a block-wise linear transition matrix to regularize the shared structure across objects. The learned dynamics can quickly adapt to new environments of unknown physical parameters and produce control signals to achieve a specified goal. Our experiments on manipulating ropes and controlling soft robots show that the proposed method has better efficiency and generalization ability than existing baselines.
****

Bridging Mode Connectivity in Loss Landscapes and Adversarial Robustness
Pu Zhao,Pin-Yu Chen,Payel Das,Karthikeyan Natesan Ramamurthy,Xue Lin
Mode connectivity provides novel geometric insights on analyzing loss landscapes and enables building high-accuracy pathways between well-trained neural networks. In this work, we propose to employ mode connectivity in loss landscapes to study the adversarial robustness of deep neural networks, and provide novel methods for improving this robustness. Our experiments cover various types of adversarial attacks applied to different network architectures and datasets. When network models are tampered with backdoor or error-injection attacks, our results demonstrate that the path connection learned using limited amount of bonafide data can effectively mitigate adversarial effects while maintaining the original accuracy on clean data. Therefore, mode connectivity provides users with the power to repair backdoored or error-injected models. We also use mode connectivity to investigate the loss landscapes of regular and robust models against evasion attacks. Experiments show that there exists a barrier in adversarial robustness loss on the path connecting regular and adversarially-trained models. A high correlation is observed between the adversarial robustness loss and the largest eigenvalue of the input Hessian matrix, for which theoretical justifications are provided. Our results suggest that mode connectivity offers a holistic tool and practical means for evaluating and improving adversarial robustness.
****

Confidence-Calibrated Adversarial Training: Towards Robust Models Generalizing Beyond the Attack Used During Training
David Stutz,Matthias Hein,Bernt Schiele
Adversarial training is the standard to train models robust against adversarial examples. However, especially for complex datasets, adversarial training incurs a significant loss in accuracy and is known to generalize poorly to stronger attacks, e.g., larger perturbations or other threat models. In this paper, we introduce confidence-calibrated adversarial training (CCAT) where the key idea is to enforce that the confidence on adversarial examples decays with their distance to the attacked examples. We show that CCAT preserves better the accuracy of normal training while robustness against adversarial examples is achieved via confidence thresholding. Most importantly, in strong contrast to adversarial training, the robustness of CCAT generalizes to larger perturbations and other threat models, not encountered during training. We also discuss our extensive work to design strong adaptive attacks against CCAT and standard adversarial training which is of independent interest. We present experimental results on MNIST, SVHN and Cifar10.
****

All SMILES Variational Autoencoder for Molecular Property Prediction and Optimization
Zaccary Alperstein,Artem Cherkasov,Jason Rolfe
Variational autoencoders (VAEs) defined over SMILES string and graph-based representations of molecules promise to improve the optimization of molecular properties, thereby revolutionizing the pharmaceuticals and materials industries. However, these VAEs are hindered by the non-unique nature of SMILES strings and the computational cost of graph convolutions. To efficiently pass messages along all paths through the molecular graph, we encode multiple SMILES strings of a single molecule using a set of stacked recurrent neural networks, harmonizing hidden r

epresentations of each atom between SMILES representations, and use attentional pooling to build a final fixed-length latent representation. By then decoding to a disjoint set of SMILES strings of the molecule, our All SMILES VAE learns an almost bijective mapping between molecules and latent representations near the high-probability-mass subspace of the prior. Our SMILES-derived but molecule-based latent representations significantly surpass the state-of-the-art in a variety of fully- and semi-supervised property regression and molecular property optimization tasks.
****

Generating Dialogue Responses From A Semantic Latent Space
Wei-Jen Ko,Avik Ray,Yilin Shen,Hongxia Jin
Generic responses are a known issue for open-domain dialog generation. Most current approaches model this one-to-many task as a one-to-one task, hence being unable to integrate information from multiple semantically similar valid responses of a prompt. We propose a novel dialog generation model that learns a semantic latent space, on which representations of semantically related sentences are close to each other. This latent space is learned by maximizing correlation between the features extracted from prompt and responses. Learning the pair relationship between the prompts and responses as a regression task on the latent space, instead of classification on the vocabulary using MLE loss, enables our model to view semantically related responses collectively. An additional autoencoder is trained, for recovering the full sentence from the latent space. Experimental results show that our proposed model eliminates the generic response problem, while achieving comparable or better coherence compared to baselines.
****

Is There Mode Collapse? A Case Study on Face Generation and Its Black-box Calibration
Zhenyu Wu,Ye Yuan,Zhaowen Wang,Jianming Zhang,Zhangyang Wang,Hailin Jin
Generative adversarial networks (GANs) nowadays are capable of producing im-ages of incredible realism. One concern raised is whether the state-of-the-artGAN's learned distribution still suffers from mode collapse. Existing evaluation metrics for image synthesis focus on low-level perceptual quality. Diversity tests of samples from GANs are usually conducted qualitatively on a small scale. In this work, we devise a set of statistical tools, that are broadly applicable to quantitatively measuring the mode collapse of GANs. Strikingly, we consistently observe strong mode collapse on several state-of-the-art GANs using our toolset. We analyze possible causes, and for the first time present two simple yet effective "black-box" methods to calibrate the GAN learned distribution, without accessing either model parameters or the original training data.
****

Overlearning Reveals Sensitive Attributes
Congzheng Song,Vitaly Shmatikov
``"Overlearning'' means that a model trained for a seemingly simple objective implicitly learns to recognize attributes and concepts that are (1) not part of the learning objective, and (2) sensitive from a privacy or bias perspective. For example, a binary gender classifier of facial images also learns to recognize races, even races that are not represented in the training data, and identities.

We demonstrate overlearning in several vision and NLP models and analyze its harmful consequences. First, inference-time representations of an overlearned model reveal sensitive attributes of the input, breaking privacy protections such as model partitioning. Second, an overlearned model can be "`re-purposed'' for a different, privacy-violating task even in the absence of the original training data.

We show that overlearning is intrinsic for some tasks and cannot be prevented by censoring unwanted attributes. Finally, we investigate where, when, and why overlearning happens during model training.
****

Gaussian MRF Covariance Modeling for Efficient Black-Box Adversarial Attacks
Anit Kumar Sahu,J. Zico Kolter,Satya Narayan Shukla
We study the problem of generating adversarial examples in a black-box setting, where we only have access to a zeroth order oracle, providing us with loss function evaluations. We employ Markov Random Fields (MRF) to exploit the structure of input data to systematically model the covariance structure of the gradients. The MRF structure in addition to Bayesian inference for the gradients facilitates one-step attacks akin to Fast Gradient Sign Method (FGSM) albeit in the black-box setting. The resulting method uses fewer queries than the current state of the art to achieve comparable performance. In particular, in the regime of lower query budgets, we show that our method is particularly effective in terms of fewer average queries with high attack accuracy while employing one-step attacks.
****
A Kolmogorov Complexity Approach to Generalization in Deep Learning
Hazar Yueksel,Kush R. Varshney,Brian Kingsbury
Deep artificial neural networks can achieve an extremely small difference between training and test accuracies on identically distributed training and test sets, which is a standard measure of generalization. However, the training and test sets may not be sufficiently representative of the empirical sample set, which consists of real-world input samples. When samples are drawn from an underrepresented or unrepresented subset during inference, the gap between the training and inference accuracies can be significant. To address this problem, we first reformulate a classification algorithm as a procedure for searching for a source code that maps input features to classes. We then derive a necessary and sufficient condition for generalization using a universal cognitive similarity metric, namely information distance, based on Kolmogorov complexity. Using this condition, we formulate an optimization problem to learn a more general classification function. To achieve this end, we extend the input features by concatenating encodings of them, and then train the classifier on the extended features. As an illustration of this idea, we focus on image classification, where we use channel codes on the input features as a systematic way to improve the degree to which the training and test sets are representative of the empirical sample set. To showcase our theoretical findings, considering that corrupted or perturbed input features belong to the empirical sample set, but typically not to the training and test sets, we demonstrate through extensive systematic experiments that, as a result of learning a more general classification function, a model trained on encoded input features is significantly more robust to common corruptions, e.g., Gaussian and shot noise, as well as adversarial perturbations, e.g., those found via projected gradient descent, than the model trained on uncoded input features.
****
Towards Modular Algorithm Induction
Daniel A. Abolafia,Rishabh Singh,Manzil Zaheer,Charles Sutton
We present a modular neural network architecture MAIN that learns algorithms given a set of input-output examples. MAIN consists of a neural controller that interacts with a variable-length input tape and learns to compose modules together with their corresponding argument choices. Unlike previous approaches, MAIN uses a general domain-agnostic mechanism for selection of modules and their arguments. It uses a general input tape layout together with a parallel history tape to indicate most recently used locations. Finally, it uses a memoryless controller with a length-invariant self-attention based input tape encoding to allow for random access to tape locations. The MAIN architecture is trained end-to-end using reinforcement learning from a set of input-output examples. We evaluate MAIN on five algorithmic tasks and show that it can learn policies that generalizes perfectly to inputs of much longer lengths than the ones used for training.
****
Optimal Strategies Against Generative Attacks
Roy Mor,Erez Peterfreund,Matan Gavish,Amir Globerson
Generative neural models have improved dramatically recently. With this progress comes the risk that such models will be used to attack systems that rely on sensor data for authentication and anomaly detection. Many such learning systems ar

e installed worldwide, protecting critical infrastructure or private data agains t malfunction and cyber attacks. We formulate the scenario of such an authentica tion system facing generative impersonation attacks, characterize it from a theo retical perspective and explore its practical implications. In particular, we as k fundamental theoretical questions in learning, statistics and information theo ry: How hard is it to detect a "fake reality"? How much data does the attacker n eed to collect before it can reliably generate nominally-looking artificial data ? Are there optimal strategies for the attacker or the authenticator? We cast th e problem as a maximin game, characterize the optimal strategy for both attacker and authenticator in the general case, and provide the optimal strategies in cl osed form for the case of Gaussian source distributions. Our analysis reveals th e structure of the optimal attack and the relative importance of data collection for both authenticator and attacker. Based on these insights we design practica l learning approaches and show that they result in models that are more robust t o various attacks on real-world data.

****

Stein Self-Repulsive Dynamics: Benefits from Past Samples

Mao Ye,Tongzheng Ren,Qiang Liu

We propose a new Stein self-repulsive dynamics for obtaining diversified samples from intractable un-normalized distributions. Our idea is to introduce Stein va riational gradient as a repulsive force to push the samples of Langevin dynamics away from the past trajectories. This simple idea allows us to significantly dec rease the auto-correlation in Langevin dynamics and hence increase the effective sample size. Importantly, as we establish in our theoretical analysis, the asym ptotic stationary distribution remains correct even with the addition of the rep ulsive force, thanks to the special properties of the Stein variational gradient . We perform extensive empirical studies of our new algorithm, showing that our method yields much higher sample efficiency and better uncertainty estimation th an vanilla Langevin dynamics.

****

Adversarially robust transfer learning

Ali Shafahi,Parsa Saadatpanah,Chen Zhu,Amin Ghiasi,Christoph Studer,David Jacobs ,Tom Goldstein

Transfer learning, in which a network is trained on one task and re-purposed on another, is often used to produce neural network classifiers when data is scarce or full-scale training is too costly. When the goal is to produce a model that is not only accurate but also adversarially robust, data scarcity and computati onal limitations become even more cumbersome.

We consider robust transfer learning, in which we transfer not only performance but also robustness from a source model to a target domain. We start by observi ng that robust networks contain robust feature extractors. By training classifie rs on top of these feature extractors, we produce new models that inherit the ro bustness of their parent networks. We then consider the case of "fine tuning" a network by re-training end-to-end in the target domain. When using lifelong lear ning strategies, this process preserves the robustness of the source network whi le achieving high accuracy. By using such strategies, it is possible to produce accurate and robust models with little data, and without the cost of adversarial training. Additionally, we can improve the generalization of adversarially trai ned models, while maintaining their robustness.

****

Doubly Robust Bias Reduction in Infinite Horizon Off-Policy Estimation

Ziyang Tang*,Yihao Feng*,Lihong Li,Dengyong Zhou,Qiang Liu

Infinite horizon off-policy policy evaluation is a highly challenging task due t o the excessively large variance of typical importance sampling (IS) estimators. Recently, Liu et al. (2018) proposed an approach that significantly reduces the variance of infinite-horizon off-policy evaluation by estimating the stationary density ratio, but at the cost of introducing potentially high risks due to the error in density ratio estimation. In this paper, we develop a bias-reduced aug mentation of their method, which can take advantage of a learned value function

to obtain higher accuracy. Our method is doubly robust in that the bias vanishes when either the density ratio or value function estimation is perfect. In general, when either of them is accurate, the bias can also be reduced. Both theoretical and empirical results show that our method yields significant advantages over previous methods.
****

Promoting Coordination through Policy Regularization in Multi-Agent Deep Reinforcement Learning

Paul Barde,Julien Roy,Félix G. Harvey,Derek Nowrouzezahrai,Christopher Pal

A central challenge in multi-agent reinforcement learning is the induction of coordination between agents of a team. In this work, we investigate how to promote inter-agent coordination using policy regularization and discuss two possible avenues respectively based on inter-agent modelling and synchronized sub-policy selection. We test each approach in four challenging continuous control tasks with sparse rewards and compare them against three baselines including MADDPG, a state-of-the-art multi-agent reinforcement learning algorithm. To ensure a fair comparison, we rely on a thorough hyper-parameter selection and training methodology that allows a fixed hyper-parameter search budget for each algorithm and environment. We consequently assess both the hyper-parameter sensitivity, sample-efficiency and asymptotic performance of each learning method. Our experiments show that the proposed methods lead to significant improvements on cooperative problems. We further analyse the effects of the proposed regularizations on the behaviors learned by the agents.
****

Contextual Text Style Transfer

Yu Cheng,Zhe Gan,Yizhe Zhang,Oussama Elachqar,Dianqi Li,Jingjing Liu

In this paper, we introduce a new task, Contextual Text Style Transfer, to translate a sentence within a paragraph context into the desired style (e.g., informal to formal, offensive to non-offensive). Two new datasets, Enron-Context and Reddit-Context, are introduced for this new task, focusing on formality and offensiveness, respectively. Two key challenges exist in contextual text style transfer: 1) how to preserve the semantic meaning of the target sentence and its consistency with the surrounding context when generating an alternative sentence with a specific style; 2) how to deal with the lack of labeled parallel data. To address these challenges, we propose a Context-Aware Style Transfer (CAST) model, which leverages both parallel and non-parallel data for joint model training. For parallel training data, CAST uses two separate encoders to encode each input sentence and its surrounding context, respectively. The encoded feature vector, together with the target style information, are then used to generate the target sentence. A classifier is further used to ensure contextual consistency of the generated sentence. In order to lever-age massive non-parallel corpus and to enhance sentence encoder and decoder training, additional self-reconstruction and back-translation losses are introduced. Experimental results on Enron-Context and Reddit-Context demonstrate the effectiveness of the proposed model over state-of-the-art style transfer methods, across style accuracy, content preservation, and contextual consistency metrics.
****

Modeling question asking using neural program generation

Ziyun Wang,Brenden M. Lake

People ask questions that are far richer, more informative, and more creative than current AI systems. We propose a neural program generation framework for modeling human question asking, which represents questions as formal programs and generates programs with an encoder-decoder based deep neural network. From extensive experiments using an information-search game, we show that our method can ask optimal questions in synthetic settings, and predict which questions humans are likely to ask in unconstrained settings. We also propose a novel grammar-based question generation framework trained with reinforcement learning, which is able to generate creative questions without supervised data.
****

Learning to Link

Maria-Florina Balcan,Travis Dick,Manuel Lang
Clustering is an important part of many modern data analysis pipelines, including network analysis and data retrieval. There are many different clustering algorithms developed by various communities, and it is often not clear which algorithm will give the best performance on a specific clustering task. Similarly, we often have multiple ways to measure distances between data points, and the best clustering performance might require a non-trivial combination of those metrics. In this work, we study data-driven algorithm selection and metric learning for clustering problems, where the goal is to simultaneously learn the best algorithm and metric for a specific application. The family of clustering algorithms we consider is parameterized linkage based procedures that includes single and complete linkage. The family of distance functions we learn over are convex combinations of base distance functions. We design efficient learning algorithms which receive samples from an application-specific distribution over clustering instances and learn a near-optimal distance and clustering algorithm from these classes. We also carry out a comprehensive empirical evaluation of our techniques showing that they can lead to significantly improved clustering performance on real-world datasets.
****

Adversarial Attacks on Copyright Detection Systems
Parsa Saadatpanah,Ali Shafahi,Tom Goldstein
It is well-known that many machine learning models are susceptible to adversarial attacks, in which an attacker evades a classifier by making small perturbations to inputs. This paper discusses how industrial copyright detection tools, which serve a central role on the web, are susceptible to adversarial attacks. We discuss a range of copyright detection systems, and why they are particularly vulnerable to attacks. These vulnerabilities are especially apparent for neural network based systems. As proof of concept, we describe a well-known music identification method and implement this system in the form of a neural net. We then attack this system using simple gradient methods. Adversarial music created this way successfully fools industrial systems, including the AudioTag copyright detector and YouTube's Content ID system. Our goal is to raise awareness of the threats posed by adversarial examples in this space and to highlight the importance of hardening copyright detection systems to attacks.
****

Detecting Extrapolation with Local Ensembles
David Madras,James Atwood,Alexander D'Amour
We present local ensembles, a method for detecting extrapolation at test time in a pre-trained model. We focus on underdetermination as a key component of extrapolation: we aim to detect when many possible predictions are consistent with the training data and model class. Our method uses local second-order information to approximate the variance of predictions across an ensemble of models from the same class. We compute this approximation by estimating the norm of the component of a test point's gradient that aligns with the low-curvature directions of the Hessian, and provide a tractable method for estimating this quantity. Experimentally, we show that our method is capable of detecting when a pre-trained model is extrapolating on test data, with applications to out-of-distribution detection, detecting spurious correlates, and active learning.
****

Global Relational Models of Source Code
Vincent J. Hellendoorn,Charles Sutton,Rishabh Singh,Petros Maniatis,David Bieber
Models of code can learn distributed representations of a program's syntax and semantics to predict many non-trivial properties of a program. Recent state-of-the-art models leverage highly structured representations of programs, such as trees, graphs and paths therein (e.g. data-flow relations), which are precise and abundantly available for code. This provides a strong inductive bias towards semantically meaningful relations, yielding more generalizable representations than classical sequence-based models. Unfortunately, these models primarily rely on graph-based message passing to represent relations in code, which makes them de facto local due to the high cost of message-passing steps, quite in contrast to m

odern, global sequence-based models, such as the Transformer. In this work, we b ridge this divide between global and structured models by introducing two new hy brid model families that are both global and incorporate structural bias: Graph Sandwiches, which wrap traditional (gated) graph message-passing layers in seque ntial message-passing layers; and Graph Relational Embedding Attention Transform ers (GREAT for short), which bias traditional Transformers with relational infor mation from graph edge types. By studying a popular, non-trivial program repair task, variable-misuse identification, we explore the relative merits of traditio nal and hybrid model families for code representation. Starting with a graph-ba sed model that already improves upon the prior state-of-the-art for this task by 20%, we show that our proposed hybrid models improve an additional 10-15%, whil e training both faster and using fewer parameters.
****

MONET: Debiasing Graph Embeddings via the Metadata-Orthogonal Training Unit
John Palowitch,Bryan Perozzi
Are Graph Neural Networks (GNNs) fair? In many real world graphs, the formation of edges is related to certain node attributes (e.g. gender, community, reputati on). In this case, any GNN using these edges will be biased by this information, as it is encoded in the structure of the adjacency matrix itself. In this pape r, we show that when metadata is correlated with the formation of node neighborh oods, unsupervised node embedding dimensions learn this metadata. This bias impl ies an inability to control for important covariates in real-world applications, such as recommendation systems.

To solve these issues, we introduce the Metadata-Orthogonal Node Embedding Train ing (MONET) unit, a general model for debiasing embeddings of nodes in a graph. MONET achieves this by ensuring that the node embeddings are trained on a hyperp lane orthogonal to that of the node metadata. This effectively organizes unstruc tured embedding dimensions into an interpretable topology-only, metadata-only di vision with no linear interactions. We illustrate the effectiveness of MONET tho ugh our experiments on a variety of real world graphs, which shows that our meth od can learn and remove the effect of arbitrary covariates in tasks such as prev enting the leakage of political party affiliation in a blog network, and thwarti ng the gaming of embedding-based recommendation systems.
****

Selection via Proxy: Efficient Data Selection for Deep Learning
Cody Coleman,Christopher Yeh,Stephen Mussmann,Baharan Mirzasoleiman,Peter Bailis ,Percy Liang,Jure Leskovec,Matei Zaharia
Data selection methods, such as active learning and core-set selection, are usef ul tools for machine learning on large datasets. However, they can be prohibitiv ely expensive to apply in deep learning because they depend on feature represent ations that need to be learned. In this work, we show that we can greatly improv e the computational efficiency by using a small proxy model to perform data sele ction (e.g., selecting data points to label for active learning). By removing hi dden layers from the target model, using smaller architectures, and training for fewer epochs, we create proxies that are an order of magnitude faster to train. Although these small proxy models have higher error rates, we find that they em pirically provide useful signals for data selection. We evaluate this "selection via proxy" (SVP) approach on several data selection tasks across five datasets: CIFAR10, CIFAR100, ImageNet, Amazon Review Polarity, and Amazon Review Full. Fo r active learning, applying SVP can give an order of magnitude improvement in da ta selection runtime (i.e., the time it takes to repeatedly train and select poi nts) without significantly increasing the final error (often within 0.1%). For c ore-set selection on CIFAR10, proxies that are over 10× faster to train than the ir larger, more accurate targets can remove up to 50% of the data without harmin g the final accuracy of the target, leading to a 1.6× end-to-end training time i mprovement.
****

Meta Learning via Learned Loss
Sarah Bechtle,Artem Molchanov,Yevgen Chebotar,Edward Grefenstette,Ludovic Righet

ti,Gaurav Sukhatme,Franziska Meier
We present a meta-learning method for learning parametric loss functions that ca n generalize across different tasks and model architectures. We develop a pipeli ne for training such loss functions, targeted at maximizing the performance of m odel learn- ing with them. We observe that the loss landscape produced by our le arned losses significantly improves upon the original task-specific losses in bo th supervised and reinforcement learning tasks. Furthermore, we show that our me ta-learning framework is flexible enough to incorporate additional information a t meta-train time. This information shapes the learned loss function such that t he environment does not need to provide this information during meta-test time.
****

Short and Sparse Deconvolution --- A Geometric Approach
Yenson Lau,Qing Qu,Han-Wen Kuo,Pengcheng Zhou,Yuqian Zhang,John Wright
Short-and-sparse deconvolution (SaSD) is the problem of extracting localized, re curring motifs in signals with spatial or temporal structure. Variants of this p roblem arise in applications such as image deblurring, microscopy, neural spike sorting, and more. The problem is challenging in both theory and practice, as na tural optimization formulations are nonconvex. Moreover, practical deconvolution problems involve smooth motifs (kernels) whose spectra decay rapidly, resulting in poor conditioning and numerical challenges. This paper is motivated by recen t theoretical advances \citep{zhang2017global,kuo2019geometry}, which characteri ze the optimization landscape of a particular nonconvex formulation of SaSD. Thi s is used to derive a provable algorithm that exactly solves certain non-practic al instances of the SaSD problem. We leverage the key ideas from this theory (sp here constraints, data-driven initialization) to develop a practical algorithm, which performs well on data arising from a range of application areas. We highli ght key additional challenges posed by the ill-conditioning of real SaSD problem s and suggest heuristics (acceleration, continuation, reweighting) to mitigate t hem. Experiments demonstrate the performance and generality of the proposed meth od.
****

If MaxEnt RL is the Answer, What is the Question?
Benjamin Eysenbach,Sergey Levine
Experimentally, it has been observed that humans and animals often make decision s that do not maximize their expected utility, but rather choose outcomes random ly, with probability proportional to expected utility. Probability matching, as this strategy is called, is equivalent to maximum entropy reinforcement learning (MaxEnt RL). However, MaxEnt RL does not optimize expected utility. In this pap er, we formally show that MaxEnt RL does optimally solve certain classes of cont rol problems with variability in the reward function. In particular, we show (1) that MaxEnt RL can be used to solve a certain class of POMDPs, and (2) that Max Ent RL is equivalent to a two-player game where an adversary chooses the reward function. These results suggest a deeper connection between MaxEnt RL, robust co ntrol, and POMDPs, and provide insight for the types of problems for which we mi ght expect MaxEnt RL to produce effective solutions. Specifically, our results s uggest that domains with uncertainty in the task goal may be especially well-sui ted for MaxEnt RL methods.
****

Stochastic Weight Averaging in Parallel: Large-Batch Training That Generalizes W ell
Vipul Gupta,Santiago Akle Serrano,Dennis DeCoste
We propose Stochastic Weight Averaging in Parallel (SWAP), an algorithm to accel erate DNN training. Our algorithm uses large mini-batches to compute an approxim ate solution quickly and then refines it by averaging the weights of multiple mo dels computed independently and in parallel. The resulting models generalize equ ally well as those trained with small mini-batches but are produced in a substan tially shorter time. We demonstrate the reduction in training time and the good generalization performance of the resulting models on the computer vision datase ts CIFAR10, CIFAR100, and ImageNet.
****

Characterizing Missing Information in Deep Networks Using Backpropagated Gradien
ts
Gukyeong Kwon,Mohit Prabhushankar,Dogancan Temel,Ghassan AlRegib
Deep networks face challenges of ensuring their robustness against inputs that c
annot be effectively represented by information learned from training data. We a
ttribute this vulnerability to the limitations inherent to activation-based repr
esentation. To complement the learned information from activation-based represen
tation, we propose utilizing a gradient-based representation that explicitly foc
uses on missing information. In addition, we propose a directional constraint on
 the gradients as an objective during training to improve the characterization o
f missing information. To validate the effectiveness of the proposed approach, w
e compare the anomaly detection performance of gradient-based and activation-bas
ed representations. We show that the gradient-based representation outperforms t
he activation-based representation by 0.093 in CIFAR-10 and 0.361 in CURE-TSR da
tasets in terms of AUROC averaged over all classes. Also, we propose an anomaly
detection algorithm that uses the gradient-based representation, denoted as Grad
Con, and validate its performance on three benchmarking datasets. The proposed m
ethod outperforms the majority of the state-of-the-art algorithms in CIFAR-10, M
NIST, and fMNIST datasets with an average AUROC of 0.664, 0.973, and 0.934, resp
ectively.
****

Scaleable input gradient regularization for adversarial robustness
Chris Finlay,Adam M Oberman
In this work we revisit gradient regularization for adversarial robustness with
some new ingredients. First, we derive new per-image theoretical robustness bou
nds based on local gradient information. These bounds strongly motivate input gr
adient regularization. Second, we implement a scaleable version of input gradie
nt regularization which avoids double backpropagation: adversarially robust Imag
eNet models are trained in 33 hours on four consumer grade GPUs. Finally, we sh
ow experimentally and through theoretical certification that input gradient regu
larization is competitive with adversarial training. Moreover we demonstrate tha
t gradient regularization does not lead to gradient obfuscation or gradient mask
ing.
****

Adjustable Real-time Style Transfer
Mohammad Babaeizadeh,Golnaz Ghiasi
Artistic style transfer is the problem of synthesizing an image with content sim
ilar to a given image and style similar to another. Although recent feed-forward
 neural networks can generate stylized images in real-time, these models produce
 a single stylization given a pair of style/content images, and the user doesn't
 have control over the synthesized output. Moreover, the style transfer depends
on the hyper-parameters of the model with varying ``optimum" for different input
 images. Therefore, if the stylized output is not appealing to the user, she/he
has to try multiple models or retrain one with different hyper-parameters to get
 a favorite stylization. In this paper, we address these issues by proposing a n
ovel method which allows adjustment of crucial hyper-parameters, after the train
ing and in real-time, through a set of manually adjustable parameters. These par
ameters enable the user to modify the synthesized outputs from the same pair of
style/content images, in search of a favorite stylized image. Our quantitative a
nd qualitative experiments indicate how adjusting these parameters is comparable
 to retraining the model with different hyper-parameters. We also demonstrate ho
w these parameters can be randomized to generate results which are diverse but s
till very similar in style and content.
****

Unsupervised Progressive Learning and the STAM Architecture
James Smith,Constantine Dovrolis
We first pose the Unsupervised Progressive Learning (UPL) problem: lear
ning salient representations from a non-stationary stream of unlabeled data in w
hich the number of object classes increases with time. If some limited labeled
data is also available, those representations can be associated with specific cl

asses, thus enabling classification tasks. To solve the UPL problem, we propose an architecture that involves an online clustering module, called Self-Taught Associative Memory (STAM). Layered hierarchies of STAM modules learn based on a combination of online clustering, novelty detection, forgetting outliers, and storing only prototypical representations rather than specific examples. The goal of this paper is to introduce the UPL problem, describe the STAM architecture, and evaluate the latter in the UPL context.
****

Wasserstein Robust Reinforcement Learning
Mohammed Amin Abdullah,Hang Ren,Haitham Bou-Ammar,Vladimir Milenkovic,Rui Luo,Mingtian Zhang,Jun Wang
Reinforcement learning algorithms, though successful, tend to over-fit to training environments, thereby hampering their application to the real-world. This paper proposes $\text{W}\text{R}^{2}\text{L}$ -- a robust reinforcement learning algorithm with significant robust performance on low and high-dimensional control tasks. Our method formalises robust reinforcement learning as a novel min-max game with a Wasserstein constraint for a correct and convergent solver. Apart from the formulation, we also propose an efficient and scalable solver following a novel zero-order optimisation method that we believe can be useful to numerical optimisation in general.
We empirically demonstrate significant gains compared to standard and robust state-of-the-art algorithms on high-dimensional MuJuCo environments
****

Knowledge Hypergraphs: Prediction Beyond Binary Relations
Bahare Fatemi,Perouz Taslakian,David Vazquez,David Poole
A Knowledge Hypergraph is a knowledge base where relations are defined on two or more entities. In this work, we introduce two embedding-based models that perform link prediction in knowledge hypergraphs:
(1) HSimplE is a shift-based method that is inspired by an existing model operating on knowledge graphs, in which the representation of an entity is a function of its position in the relation, and (2) HypE is a convolution-based method which disentangles the representation of an entity from its position in the relation. We test our models on two new knowledge hypergraph datasets that we obtain from Freebase, and show that both HSimplE and HypE are more effective in predicting links in knowledge hypergraphs than the proposed baselines and existing methods.
Our experiments show that HypE outperforms HSimplE when trained with fewer parameters and when tested on samples that contain at least one entity in a position never encountered during training.
****

Dynamics-Aware Unsupervised Discovery of Skills
Archit Sharma,Shixiang Gu,Sergey Levine,Vikash Kumar,Karol Hausman
Conventionally, model-based reinforcement learning (MBRL) aims to learn a global model for the dynamics of the environment. A good model can potentially enable planning algorithms to generate a large variety of behaviors and solve diverse tasks. However, learning an accurate model for complex dynamical systems is difficult, and even then, the model might not generalize well outside the distribution of states on which it was trained. In this work, we combine model-based learning with model-free learning of primitives that make model-based planning easy. To that end, we aim to answer the question: how can we discover skills whose outcomes are easy to predict? We propose an unsupervised learning algorithm, Dynamics-Aware Discovery of Skills (DADS), which simultaneously discovers predictable behaviors and learns their dynamics. Our method can leverage continuous skill spaces, theoretically, allowing us to learn infinitely many behaviors even for high-dimensional state-spaces. We demonstrate that zero-shot planning in the learned latent space significantly outperforms standard MBRL and model-free goal-conditioned RL, can handle sparse-reward tasks, and substantially improves over prior hierarchical RL methods for unsupervised skill discovery.
****

A Fine-Grained Spectral Perspective on Neural Networks

Greg Yang,Hadi Salman
Are neural networks biased toward simple functions?
Does depth always help learn more complex features?
Is training the last layer of a network as good as training all layers?
These questions seem unrelated at face value, but in this work we give all of th em a common treatment from the spectral perspective.
We will study the spectra of the *Conjugate Kernel, CK,* (also called the *Neura l Network-Gaussian Process Kernel*), and the *Neural Tangent Kernel, NTK*.
Roughly, the CK and the NTK tell us respectively ``"what a network looks like at initialization" and "``what a network looks like during and after training."
Their spectra then encode valuable information about the initial distribution an d the training and generalization properties of neural networks.
By analyzing the eigenvalues, we lend novel insights into the questions put fort h at the beginning, and we verify these insights by extensive experiments of neu ral networks.
We believe the computational tools we develop here for analyzing the spectra of CK and NTK serve as a solid foundation for future studies of deep neural network s.
We have open-sourced the code for it and for generating the plots in this paper at github.com/jxVmnLgedVwv6mNcGCBy/NNspectra.
****

Energy-Aware Neural Architecture Optimization with Fast Splitting Steepest Desce nt
Dilin Wang,Meng Li,Lemeng Wu,Vikas Chandra,Qiang Liu
Designing energy-efficient networks is of critical importance for enabling state -of-the-art deep learning in mobile and edge settings where the computation and energy budgets are highly limited. Recently, Wu et al. (2019) framed the search of efficient neural architectures into a continuous splitting process: it iterat ively splits existing neurons into multiple off-springs to achieve progressive l oss minimization, thus finding novel architectures by gradually growing the neu ral network. However, this method was not specifically tailored for designing en ergy-efficient networks, and is computationally expensive on large-scale benchma rks. In this work, we substantially improve Wu et al. (2019) in two significant ways: 1) we incorporate the energy cost of splitting different neurons to bett er guide the splitting process, thereby discovering more energy-efficient networ k architectures; 2) we substantially speed up the splitting process of Wu et al. (2019), which requires expensive eigen-decomposition, by proposing a highly sca lable Rayleigh-quotient stochastic gradient algorithm. Our fast algorithm allow s us to reduce the computational cost of splitting to the same level of typical back-propagation updates and enables efficient implementation on GPU. Extensive empirical results show that our method can train highly accurate and energy-effi cient networks on challenging datasets such as ImageNet, improving a variety of baselines, including the pruning-based methods and expert-designed architectur es.
****

Unpaired Point Cloud Completion on Real Scans using Adversarial Training
Xuelin Chen,Baoquan Chen,Niloy J. Mitra
As 3D scanning solutions become increasingly popular, several deep learning setu ps have been developed for the task of scan completion, i.e., plausibly filling in regions that were missed in the raw scans. These methods, however, largely re ly on supervision in the form of paired training data, i.e., partial scans with corresponding desired completed scans. While these methods have been successfull y demonstrated on synthetic data, the approaches cannot be directly used on real scans in absence of suitable paired training data. We develop a first approach that works directly on input point clouds, does not require paired training data , and hence can directly be applied to real scans for scan completion. We evalu ate the approach qualitatively on several real-world datasets (ScanNet, Matterpo rt3D, KITTI), quantitatively on 3D-EPN shape completion benchmark dataset, and d emonstrate realistic completions under varying levels of incompleteness.

****
Efficient Riemannian Optimization on the Stiefel Manifold via the Cayley Transfo
rm
Jun Li,Fuxin Li,Sinisa Todorovic
Strictly enforcing orthonormality constraints on parameter matrices has been sho
wn advantageous in deep learning. This amounts to Riemannian optimization on the
 Stiefel manifold, which, however, is computationally expensive. To address this
 challenge, we present two main contributions: (1) A new efficient retraction ma
p based on an iterative Cayley transform for optimization updates, and (2) An im
plicit vector transport mechanism based on the combination of a projection of th
e momentum and the Cayley transform on the Stiefel manifold. We specify two new
optimization algorithms: Cayley SGD with momentum, and Cayley ADAM on the Stiefe
l manifold. Convergence of Cayley SGD is theoretically analyzed. Our experiments
 for CNN training demonstrate that both algorithms: (a) Use less running time pe
r iteration relative to existing approaches that enforce orthonormality of CNN p
arameters; and (b) Achieve faster convergence rates than the baseline SGD and AD
AM algorithms without compromising the performance of the CNN. Cayley SGD and Ca
yley ADAM are also shown to reduce the training time for optimizing the unitary
transition matrices in RNNs.
****
DIME: AN INFORMATION-THEORETIC DIFFICULTY MEASURE FOR AI DATASETS
Peiliang Zhang,Huan Wang,Nikhil Naik,Caiming Xiong,Richard Socher
Evaluating the relative difficulty of widely-used benchmark datasets across time
 and across data modalities is important for accurately measuring progress in ma
chine learning. To help tackle this problem, we proposeDIME, an information-the
oretic DIfficulty MEasure for datasets, based on conditional entropy estimation
of the sample-label distribution. Theoretically, we prove a model-agnostic and
 modality-agnostic lower bound on the 0-1 error by extending Fano’s inequality t
o the common supervised learning scenario where labels are discrete and features
 are continuous. Empirically, we estimate this lower bound using a neural networ
k to compute DIME. DIME can be decomposed into components attributable to the da
ta distribution and the number of samples. DIME can also compute per-class diff
iculty scores. Through extensive experiments on both vision and language dataset
s, we show that DIME is well-aligned with empirically observed performance of st
ate-of-the-art machine learning models. We hope that DIME can aid future dataset
 design and model-training strategies.
****
Structured consistency loss for semi-supervised semantic segmentation
JongMok Kim,Joo Young Jang,Hyunwoo Park
The consistency loss has played a key role in solving problems in recent studies
 on semi-supervised learning. Yet extant studies with the consistency loss are l
imited to its application to classification tasks; extant studies on semi-superv
ised semantic segmentation rely on pixel-wise classification, which does not ref
lect the structured nature of characteristics in prediction. We propose a struct
ured consistency loss to address this limitation of extant studies. Structured c
onsistency loss promotes consistency in inter-pixel similarity between teacher a
nd student networks. Specifically, collaboration with CutMix optimizes the effic
ient performance of semi-supervised semantic segmentation with structured consis
tency loss by reducing computational burden dramatically. The superiority of pro
posed method is verified with the Cityscapes; The Cityscapes benchmark results w
ith validation and with test data are 81.9 mIoU and 83.84 mIoU respectively. Thi
s ranks the first place on the pixel-level semantic labeling task of Cityscapes
benchmark suite. To the best of our knowledge, we are the first to present the s
uperiority of state-of-the-art semi-supervised learning in semantic segmentation
.
****
AMRL: Aggregated Memory For Reinforcement Learning
Jacob Beck,Kamil Ciosek,Sam Devlin,Sebastian Tschiatschek,Cheng Zhang,Katja Hofm
ann
In many partially observable scenarios, Reinforcement Learning (RL) agents must

rely on long-term memory in order to learn an optimal policy. We demonstrate tha t using techniques from NLP and supervised learning fails at RL tasks due to sto chasticity from the environment and from exploration. Utilizing our insights on the limitations of traditional memory methods in RL, we propose AMRL, a class of models that can learn better policies with greater sample efficiency and are re silient to noisy inputs. Specifically, our models use a standard memory module t o summarize short-term context, and then aggregate all prior states from the sta ndard model without respect to order. We show that this provides advantages both in terms of gradient decay and signal-to-noise ratio over time. Evaluating in M inecraft and maze environments that test long-term memory, we find that our mode l improves average return by 19% over a baseline that has the same number of par ameters and by 9% over a stronger baseline that has far more parameters.
****

Adapting Behaviour for Learning Progress
Tom Schaul,Diana Borsa,David Ding,David Szepesvari,Georg Ostrovski,Will Dabney,S imon Osindero
Determining what experience to generate to best facilitate learning (i.e. explor ation) is one of the distinguishing features and open challenges in reinforcemen t learning. The advent of distributed agents that interact with parallel instanc es of the environment has enabled larger scale and greater flexibility, but has not removed the need to tune or tailor exploration to the task, because the idea l data for the learning algorithm necessarily depends on its process of learning . We propose to dynamically adapt the data generation by using a non-stationary multi-armed bandit to optimize a proxy of the learning progress. The data distri bution is controlled via modulating multiple parameters of the policy (such as s tochasticity, consistency or optimism) without significant overhead. The adaptat ion speed of the bandit can be increased by exploiting the factored modulation s tructure. We demonstrate on a suite of Atari 2600 games how this unified approac h produces results comparable to per-task tuning at a fraction of the cost.
****

Pretraining boosts out-of-domain robustness for pose estimation
Alexander Mathis,Mert Yüksekgönül,Byron Rogers,Matthias Bethge,Mackenzie W. Math is
Deep neural networks are highly effective tools for human and animal pose estima tion. However, robustness to out-of-domain data remains a challenge. Here, we pr obe the transfer and generalization ability for pose estimation with two archite cture classes (MobileNetV2s and ResNets) pretrained on ImageNet. We generated a novel dataset of 30 horses that allowed for both within-domain and out-of-domain (unseen horse) testing. We find that pretraining on ImageNet strongly improves out-of-domain performance. Moreover, we show that for both pretrained and networ ks trained from scratch, better ImageNet-performing architectures perform better for pose estimation, with a substantial improvement on out-of-domain data when pretrained. Collectively, our results demonstrate that transfer learning is part icularly beneficial for out-of-domain robustness.
****

GraphMix: Regularized Training of Graph Neural Networks for Semi-Supervised Lear ning
Vikas Verma,Meng Qu,Alex Lamb,Yoshua Bengio,Juho Kannala,Jian Tang
We present GraphMix, a regularization technique for Graph Neural Network based s emi-supervised object classification, leveraging the recent advances in the regu larization of classical deep neural networks. Specifically, we propose a unified approach in which we train a fully-connected network jointly with the graph neu ral network via parameter sharing, interpolation-based regularization and self-p redicted-targets. Our proposed method is architecture agnostic in the sense that it can be applied to any variant of graph neural networks which applies a param etric transformation to the features of the graph nodes. Despite its simplicity, with GraphMix we can consistently improve results and achieve or closely match state-of-the-art performance using even simpler architectures such as Graph Conv olutional Networks, across three established graph benchmarks: Cora, Citeseer an d Pubmed citation network datasets, as well as three newly proposed datasets : C

ora-Full, Co-author-CS and Co-author-Physics.
****
Synthetic vs Real: Deep Learning on Controlled Noise
Lu Jiang,Di Huang,Weilong Yang
Performing controlled experiments on noisy data is essential in thoroughly understanding deep learning across a spectrum of noise levels. Due to the lack of suitable datasets, previous research have only examined deep learning on controlled synthetic noise, and real-world noise has never been systematically studied in a controlled setting. To this end, this paper establishes a benchmark of real-world noisy labels at 10 controlled noise levels. As real-world noise possesses unique properties, to understand the difference, we conduct a large-scale study across a variety of noise levels and types, architectures, methods, and training settings. Our study shows that: (1) Deep Neural Networks (DNNs) generalize much better on real-world noise. (2) DNNs may not learn patterns first on real-world noisy data. (3) When networks are fine-tuned, ImageNet architectures generalize well on noisy data. (4) Real-world noise appears to be less harmful, yet it is more difficult for robust DNN methods to improve. (5) Robust learning methods that work well on synthetic noise may not work as well on real-world noise, and vice versa. We hope our benchmark, as well as our findings, will facilitate deep learning research on noisy data.

****
Detecting malicious PDF using CNN
Raphael Fettaya,Yishay Mansour
Malicious PDF files represent one of the biggest threats to computer security. To
detect them, significant research has been done using handwritten signatures or machine learning based on manual feature extraction. Those approaches are both time-consuming, requires significant prior knowledge and the list of features has
to be updated with each newly discovered vulnerability. In this work, we propose a novel algorithm that uses a Convolutional Neural Network (CNN) on the byte level of the file, without any handcrafted features. We show, using a data set of 130000 files, that our approach maintains a high detection rate (96%) of PDF malware and even detects new malicious files, still undetected by most antiviruses.
Using automatically generated features from our CNN network, and applying a clustering algorithm, we also obtain high similarity between the antiviruses’ labels
and the resulting clusters.
****
NESTED LEARNING FOR MULTI-GRANULAR TASKS
Raphaël Achddou,J. Matias Di Martino,Guillermo Sapiro
Standard deep neural networks (DNNs) used for classification are trained in an end-to-end fashion for very specific tasks - object recognition, face identification, character recognition, etc. This specificity often leads to overconfident models that generalize poorly to samples that are not from the original training distribution. Moreover, they do not allow to leverage information from heterogeneously annotated data, where for example, labels may be provided with different levels of granularity. Finally, standard DNNs do not produce results with simultaneous different levels of confidence for different levels of detail, they are most commonly an all or nothing approach. To address these challenges, we introduce the problem of nested learning: how to obtain a hierarchical representation of the input such that a coarse label can be extracted first, and sequentially refine this representation to obtain successively refined predictions, all of them with the corresponding confidence. We explicitly enforce this behaviour by creating a sequence of nested information bottlenecks. Looking at the problem of nested learning from an information theory perspective, we design a network topology with two important properties. First, a sequence of low dimensional (nested) feature embeddings are enforced. Then we show how the explicit combination of ne

sted outputs can improve both robustness and finer predictions. Experimental res
ults on CIFAR-10, MNIST, and FASHION-MNIST demonstrate that nested learning outp
erforms the same network trained in the standard end-to-end fashion. Since the n
etwork can be naturally trained with mixed data labeled at different levels of n
ested details, we also study what is the most efficient way of annotating data,
when a fixed training budget is given and the cost of labels increases with the
levels in the nested hierarchy.
****
Scalable Model Compression by Entropy Penalized Reparameterization
Deniz Oktay,Johannes Ballé,Saurabh Singh,Abhinav Shrivastava
We describe a simple and general neural network weight compression approach, in
which the network parameters (weights and biases) are represented in a “latent”
space, amounting to a reparameterization. This space is equipped with a learned
probability model, which is used to impose an entropy penalty on the parameter r
epresentation during training, and to compress the representation using a simple
 arithmetic coder after training. Classification accuracy and model compressibil
ity is maximized jointly, with the bitrate–accuracy trade-off specified by a hyp
erparameter. We evaluate the method on the MNIST, CIFAR-10 and ImageNet classifi
cation benchmarks using six distinct model architectures. Our results show that
state-of-the-art model compression can be achieved in a scalable and general way
 without requiring complex procedures such as multi-stage training.
****
Dynamic Time Lag Regression: Predicting What & When
Mandar Chandorkar,Cyril Furtlehner,Bala Poduval,Enrico Camporeale,Michele Sebag
This paper tackles a new regression problem, called Dynamic Time-Lag Regression
(DTLR), where a cause signal drives an effect signal with an unknown time delay.
The motivating application, pertaining to space weather modelling, aims to predi
ct the near-Earth solar wind speed based on estimates of the Sun's coronal magn
etic field.
DTLR differs from mainstream regression and from sequence-to-sequence learning i
n two respects: firstly, no ground truth (e.g., pairs of associated sub-sequence
s) is available; secondly, the cause signal contains much information irrelevant
 to the effect signal (the solar magnetic field governs the solar wind propagati
on in the heliosphere, of which the Earth's magnetosphere is but a minuscule reg
ion).

A Bayesian approach is presented to tackle the specifics of the DTLR problem, wi
th theoretical justifications based on linear stability analysis. A proof of con
cept on synthetic problems is presented. Finally, the empirical results on the s
olar wind modelling task improve on the state of the art in solar wind forecasti
ng.
****
On summarized validation curves and generalization
Mohammad Hashir,Yoshua Bengio,Joseph Paul Cohen
The validation curve is widely used for model selection and hyper-parameter sear
ch with the curve usually summarized over all the training tasks. However, this
summarization tends to lose the intricacies of the per-task curves and it isn't
able to reflect if all the tasks are at their validation optimum even if the sum
marized curve might be. In this work, we explore this loss of information, how i
t affects the model at testing and how to detect it using interval plots. We pro
pose two techniques as a proof-of-concept of the potential gain in the test perf
ormance when per-task validation curves are accounted for. Our experiments on th
ree large datasets show up to a 2.5% increase (averaged over multiple trials) in
 the test accuracy rate when model selection uses the per-task validation maximu
ms instead of the summarized validation maximum. This potential increase is not
a result of any modification to the model but rather at what point of training t
he weights were selected from. This presents an exciting direction for new train
ing and model selection techniques that rely on more than just averaged metrics.

****

Convolutional Bipartite Attractor Networks
Michael L. Iuzzolino,Yoram Singer,Michael C. Mozer
In human perception and cognition, a fundamental operation that brains perform i
s interpretation: constructing coherent neural states from noisy, incomplete, an
d intrinsically ambiguous evidence. The problem of interpretation is well matche
d to an early and often overlooked architecture, the attractor network---a recur
rent neural net that performs constraint satisfaction, imputation of missing fea
tures, and clean up of noisy data via energy minimization dynamics. We revisit a
ttractor nets in light of modern deep learning methods and propose a convolution
al bipartite architecture with a novel training loss, activation function, and c
onnectivity constraints. We tackle larger problems than have been previously exp
lored with attractor nets and demonstrate their potential for image completion a
nd super-resolution. We argue that this architecture is better motivated than ev
er-deeper feedforward models and is a viable alternative to more costly sampling
-based generative methods on a range of supervised and unsupervised tasks.
****

New Loss Functions for Fast Maximum Inner Product Search
Ruiqi Guo,Quan Geng,David Simcha,Felix Chern,Phil Sun,Sanjiv Kumar
Quantization based methods are popular for solving large scale maximum inner pro
duct search problems. However, in most traditional quantization works, the objec
tive is to minimize the reconstruction error for datapoints to be searched. In t
his work, we focus directly on minimizing error in inner product approximation a
nd derive a new class of quantization loss functions. One key aspect of the new
loss functions is that we weight the error term based on the value of the inner
product, giving more importance to pairs of queries and datapoints whose inner p
roducts are high. We provide theoretical grounding to the new quantization loss
function, which is simple, intuitive and able to work with a variety of quantiza
tion techniques, including binary quantization and product quantization. We cond
uct experiments on public benchmarking datasets \url{http://ann-benchmarks.com}
to demonstrate that our method using the new objective outperforms other state-o
f-the-art methods. We are committed to release our source code.
****

Lipschitz Lifelong Reinforcement Learning
Erwan Lecarpentier,David Abel,Kavosh Asadi,Yuu Jinnai,Emmanuel Rachelson,Michael
 L. Littman
We consider the problem of reusing prior experience when an agent is facing a se
ries of Reinforcement Learning (RL) tasks. We introduce a novel metric between M
arkov Decision Processes and focus on the study and exploitation of the optimal
value function's Lipschitz continuity in the task space with respect to that met
ric. These theoretical results lead us to a value transfer method for Lifelong R
L, which we use to build a PAC-MDP algorithm that exploits continuity to acceler
ate learning. We illustrate the benefits of the method in Lifelong RL experiment
s.
****

Local Label Propagation for Large-Scale Semi-Supervised Learning
Chengxu Zhuang,Chaofei Fan,Xuehao Ding,Divyanshu Murli,Daniel Yamins
A significant issue in training deep neural networks to solve supervised learnin
g
tasks is the need for large numbers of labeled datapoints. The goal of semisuper
vised learning is to leverage ubiquitous unlabeled data, together with small qua
ntities of labeled data, to achieve high task performance. Though substantial re
cent progress has been made in developing semi-supervised algorithms that are ef
fective for comparatively small datasets, many of these techniques do not scale
readily to the large (unlabeled) datasets characteristic of real-world applicati
ons. In this paper we introduce a novel approach to scalable semi-supervised lea
rning, called Local Label Propagation (LLP). Extending ideas from recent work on
 unsupervised embedding learning, LLP first embeds datapoints, labeled and other
wise, in a common latent space using a deep neural network. It then propagates p
seudolabels from known to unknown datapoints in a manner that depends on the loc
al geometry of the embedding, taking into account both inter-point distance and

local data density as a weighting on propagation likelihood. The parameters of t
he deep embedding are then trained to simultaneously maximize pseudolabel catego
rization performance as well as a metric of the clustering of datapoints within
each psuedo-label group, iteratively alternating stages of network training and
label propagation. We illustrate the utility of the LLP method on the ImageNet d
ataset, achieving results that outperform previous state-of-the-art scalable sem
i-supervised learning algorithms by large margins, consistently across a wide va
riety of training regimes. We also show that the feature representation learned
with LLP transfers well to scene recognition in the Places 205 dataset.
****

Improved Mutual Information Estimation
Youssef Mroueh*,Igor Melnyk*,Pierre Dognin*,Jerret Ross*,Tom Sercu*
We propose a new variational lower bound on the KL divergence and show that the
Mutual Information (MI) can be estimated by maximizing this bound using a witnes
s function on a hypothesis function class and an auxiliary scalar variable. If t
he function class is in a Reproducing Kernel Hilbert Space (RKHS), this leads to
 a jointly convex problem. We analyze the bound by deriving its dual formulation
 and show its connection to a likelihood ratio estimation problem. We show that
the auxiliary variable introduced in our variational form plays the role of a La
grange multiplier that enforces a normalization constraint on the likelihood rat
io. By extending the function space to neural networks, we propose an efficient
neural MI estimator, and validate its performance on synthetic examples, showing
 advantage over the existing baselines. We then demonstrate the strength of our
estimator in large-scale self-supervised representation learning through MI maxi
mization.
****

Semi-Supervised Generative Modeling for Controllable Speech Synthesis
Raza Habib,Soroosh Mariooryad,Matt Shannon,Eric Battenberg,RJ Skerry-Ryan,Daisy
Stanton,David Kao,Tom Bagby
We present a novel generative model that combines state-of-the-art neural text-
to-speech (TTS) with semi-supervised probabilistic latent variable models. By pr
oviding partial supervision to some of the latent variables, we are able to forc
e them to take on consistent and interpretable purposes, which previously hasn’t
 been possible with purely unsupervised methods. We demonstrate that our model i
s able to reliably discover and control important but rarely labelled attributes
 of speech, such as affect and speaking rate, with as little as 1% (30 minutes)
supervision. Even at such low supervision levels we do not observe a degradation
 of synthesis quality compared to a state-of-the-art baseline. We will release a
udio samples at https://google.github.io/tacotron/publications/semisupervised_ge
nerative_modeling_for_controllable_speech_synthesis/.
****

Towards Physics-informed Deep Learning for Turbulent Flow Prediction
Rui Wang,Karthik Kashinath,Mustafa Mustafa,Adrian Albert,Rose Yu
While deep learning has shown tremendous success in a wide range of domains, it
remains a grand challenge to incorporate physical principles in a systematic man
ner to the design, training, and inference of such models. In this paper, we aim
 to predict turbulent flow by learning its highly nonlinear dynamics from spatio
temporal velocity fields of large-scale fluid flow simulations of relevance to t
urbulence modeling and climate modeling. We adopt a hybrid approach by marrying
two well-established turbulent flow simulation techniques with deep learning. Sp
ecifically, we introduce trainable spectral filters in a coupled model of Reynol
ds-averaged Navier-Stokes (RANS) and Large Eddy Simulation (LES), followed by a
specialized U-net for prediction. Our approach, which we call Turbulent-Flow Net
 (TF-Net), is grounded in a principled physics model, yet offers the flexibility
 of learned representations. We compare our model, TF-Net, with state-of-the-art
 baselines and observe significant reductions in error for predictions 60 frames
 ahead. Most significantly, our method predicts physical fields that obey desira
ble physical characteristics, such as conservation of mass, whilst faithfully em
ulating the turbulent kinetic energy field and spectrum, which are critical for
accurate prediction of turbulent flows.

****
Neural Text Generation With Unlikelihood Training
Sean Welleck,Ilia Kulikov,Stephen Roller,Emily Dinan,Kyunghyun Cho,Jason Weston
Neural text generation is a key tool in natural language applications, but it is well known there are major problems at its core. In particular, standard likelihood training and decoding leads to dull and repetitive outputs. While some post-hoc fixes have been proposed, in particular top-k and nucleus sampling, they do not address the fact that the token-level probabilities predicted by the model are poor. In this paper we show that the likelihood objective itself is at fault, resulting in a model that assigns too much probability to sequences containing repeats and frequent words, unlike those from the human training distribution. We propose a new objective, unlikelihood training, which forces unlikely generations to be assigned lower probability by the model. We show that both token and sequence level unlikelihood training give less repetitive, less dull text while maintaining perplexity, giving superior generations using standard greedy or beam search. According to human evaluations, our approach with standard beam search also outperforms the currently popular decoding methods of nucleus sampling or beam blocking, thus providing a strong alternative to existing techniques.
****
Pure and Spurious Critical Points: a Geometric Study of Linear Networks
Matthew Trager,Kathlén Kohn,Joan Bruna
The critical locus of the loss function of a neural network is determined by the geometry of the functional space and by the parameterization of this space by the network's weights. We introduce a natural distinction between pure critical points, which only depend on the functional space, and spurious critical points, which arise from the parameterization. We apply this perspective to revisit and extend the literature on the loss function of linear neural networks. For this type of network, the functional space is either the set of all linear maps from input to output space, or a determinantal variety, i.e., a set of linear maps with bounded rank. We use geometric properties of determinantal varieties to derive new results on the landscape of linear networks with different loss functions and different parameterizations. Our analysis clearly illustrates that the absence of "bad" local minima in the loss landscape of linear networks is due to two distinct phenomena that apply in different settings: it is true for arbitrary smooth convex losses in the case of architectures that can express all linear maps ("filling architectures") but it holds only for the quadratic loss when the functional space is a determinantal variety ("non-filling architectures"). Without any assumption on the architecture, smooth convex losses may lead to landscapes with many bad minima.
****
Surrogate-Based Constrained Langevin Sampling With Applications to Optimal Material Configuration Design
Thanh V Nguyen,Youssef Mroueh,Samuel C. Hoffman,Payel Das,Pierre Dognin,Giuseppe Romano,Chinmay Hegde
We consider the problem of generating configurations that satisfy physical constraints for optimal material nano-pattern design, where multiple (and often conflicting) properties need to be simultaneously satisfied. Consider, for example, the trade-off between thermal resistance, electrical conductivity, and mechanical stability needed to design a nano-porous template with optimal thermoelectric efficiency. To that end, we leverage the posterior regularization framework and show that this constraint satisfaction problem can be formulated as sampling from a Gibbs distribution. The main challenges come from the black-box nature ofthose physical constraints, since they are obtained via solving highly non-linearPDEs. To overcome those difficulties, we introduce Surrogate-based Constrained Langevin dynamics for black-box sampling. We explore two surrogate approaches. The first approach exploits zero-order approximation of gradients in the Langevin Sampling and we refer to it as Zero-Order Langevin. In practice, this approach can be prohibitive since we still need to often query the expensive PDE solvers. The second approach approximates the gradients in the Langevin dynamics with deep neural networks, allowing us an efficient sampling strategy using the surrogate

model. We prove the convergence of those two approaches when the target distribu
tion is log-concave and smooth. We show the effectiveness of both approaches in
designing optimal nano-porous material configurations, where the goal is to prod
uce nano-pattern templates with low thermal conductivity and reasonable mechanic
al stability.
****

Learning Heuristics for Quantified Boolean Formulas through Reinforcement Learni
ng
Gil Lederman,Markus Rabe,Sanjit Seshia,Edward A. Lee
We demonstrate how to learn efficient heuristics for automated reasoning algorit
hms for quantified Boolean formulas through deep reinforcement learning. We focu
s on a backtracking search algorithm, which can already solve formulas of impres
sive size - up to hundreds of thousands of variables. The main challenge is to f
ind a representation of these formulas that lends itself to making predictions i
n a scalable way. For a family of challenging problems, we learned a heuristic t
hat solves significantly more formulas compared to the existing handwritten heur
istics.
****

Mean Field Models for Neural Networks in Teacher-student Setting
Lexing Ying,Yuandong Tian
Mean field models have provided a convenient framework for understanding the tra
ining dynamics for certain neural networks in the infinite width limit. The resu
lting mean field equation characterizes the evolution of the time-dependent empi
rical distribution of the network parameters. Following this line of work, this
paper first focuses on the teacher-student setting. For the two-layer networks,
we derive the necessary condition of the stationary distributions of the mean fi
eld equation and explain an empirical phenomenon concerning training speed diffe
rences using the Wasserstein flow description. Second, we apply this approach to
 two extended ResNet models and characterize the necessary condition of stationa
ry distributions in the teacher-student setting.
****

A Causal View on Robustness of Neural Networks
Cheng Zhang,Yingzhen Li
We present a causal view on the robustness of neural networks against input mani
pulations, which applies not only to traditional classification tasks but also t
o general measurement data. Based on this view, we design a deep causal manipula
tion augmented model (deep CAMA) which explicitly models the manipulations of da
ta as a cause to the observed effect variables. We further develop data augmenta
tion and test-time fine-tuning methods to improve deep CAMA's robustness. When c
ompared with discriminative deep neural networks, our proposed model shows super
ior robustness against unseen manipulations. As a by-product, our model achieves
 disentangled representation which separates the representation of manipulations
 from those of other latent causes.
****

Striving for Simplicity in Off-Policy Deep Reinforcement Learning
Rishabh Agarwal,Dale Schuurmans,Mohammad Norouzi
This paper advocates the use of offline (batch) reinforcement learning (RL) to h
elp (1) isolate the contributions of exploitation vs. exploration in off-policy
deep RL, (2) improve reproducibility of deep RL research, and (3) facilitate the
 design of simpler deep RL algorithms. We propose an offline RL benchmark on Ata
ri 2600 games comprising all of the replay data of a DQN agent. Using this bench
mark, we demonstrate that recent off-policy deep RL algorithms, even when traine
d solely on logged DQN data, can outperform online DQN. We present Random Ensemb
le Mixture (REM), a simple Q-learning algorithm that enforces optimal Bellman co
nsistency on random convex combinations of multiple Q-value estimates. The REM a
lgorithm outperforms more complex RL agents such as C51 and QR-DQN on the offlin
e Atari benchmark and performs comparably in the online setting.
****

White Box Network: Obtaining a right composition ordering of functions
Eun saem Lee,Hyung Ju Hwang

Neural networks have significantly benefitted real-world tasks. The universality of a neural network enables the approximation of any type of continuous functions. However, a neural network is regarded as a non-interpretable black box model, and this is fatal to reverse engineering as the main goal of reverse engineering is to reveal the structure or design of a target function instead of approximating it. Therefore, we propose a new type of a function constructing network, called the white box network. This network arranges function blocks to construct a target function to reveal its design. The network uses discretized layers, thus rendering the model interpretable without disordering the function blocks. Additionally, we introduce an end-to-end PathNet structure through this discretization by considering the function blocks as neural networks
****

Deep neuroethology of a virtual rodent
Josh Merel,Diego Aldarondo,Jesse Marshall,Yuval Tassa,Greg Wayne,Bence Olveczky
Parallel developments in neuroscience and deep learning have led to mutually productive exchanges, pushing our understanding of real and artificial neural networks in sensory and cognitive systems. However, this interaction between fields is less developed in the study of motor control. In this work, we develop a virtual rodent as a platform for the grounded study of motor activity in artificial models of embodied control. We then use this platform to study motor activity across contexts by training a model to solve four complex tasks. Using methods familiar to neuroscientists, we describe the behavioral representations and algorithms employed by different layers of the network using a neuroethological approach to characterize motor activity relative to the rodent's behavior and goals. We find that the model uses two classes of representations which respectively encode the task-specific behavioral strategies and task-invariant behavioral kinematics. These representations are reflected in the sequential activity and population dynamics of neural subpopulations. Overall, the virtual rodent facilitates grounded collaborations between deep reinforcement learning and motor neuroscience.
****

Emergence of functional and structural properties of the head direction system by optimization of recurrent neural networks
Christopher J. Cueva,Peter Y. Wang,Matthew Chin,Xue-Xin Wei
Recent work suggests goal-driven training of neural networks can be used to model neural activity in the brain. While response properties of neurons in artificial neural networks bear similarities to those in the brain, the network architectures are often constrained to be different. Here we ask if a neural network can recover both neural representations and, if the architecture is unconstrained and optimized, also the anatomical properties of neural circuits. We demonstrate this in a system where the connectivity and the functional organization have been characterized, namely, the head direction circuit of the rodent and fruit fly. We trained recurrent neural networks (RNNs) to estimate head direction through integration of angular velocity. We found that the two distinct classes of neurons observed in the head direction system, the Compass neurons and the Shifter neurons, emerged naturally in artificial neural networks as a result of training. Furthermore, connectivity analysis and in-silico neurophysiology revealed structural and mechanistic similarities between artificial networks and the head direction system. Overall, our results show that optimization of RNNs in a goal-driven task can recapitulate the structure and function of biological circuits, suggesting that artificial neural networks can be used to study the brain at the level of both neural activity and anatomical organization.
****

Causal Induction from Visual Observations for Goal Directed Tasks
Suraj Nair,Yuke Zhu,Silvio Savarese,Li Fei-Fei
Causal reasoning has been an indispensable capability for humans and other intelligent animals to interact with the physical world. In this work, we propose to endow an artificial agent with the capability of causal reasoning for completing goal-directed tasks. We develop learning-based approaches to inducing causal knowledge in the form of directed acyclic graphs, which can be used to contextualize a learned goal-conditional policy to perform tasks in novel environments with

latent causal structures. We leverage attention mechanisms in our causal induction model and goal-conditional policy, enabling us to incrementally generate the causal graph from the agent's visual observations and to selectively use the induced graph for determining actions. Our experiments show that our method effectively generalizes towards completing new tasks in novel environments with previously unseen causal structures.

****

Duration-of-Stay Storage Assignment under Uncertainty

Michael Lingzhi Li,Elliott Wolf,Daniel Wintz

Storage assignment, the act of choosing what goods are placed in what locations in a warehouse, is a central problem of supply chain logistics. Past literature has shown that the optimal method to assign pallets is to arrange them in increasing duration of stay in the warehouse (the Duration-of-Stay, or DoS, method), but the methodology requires perfect prior knowledge of DoS for each pallet, which is unknown and uncertain under realistic conditions. Attempts to predict DoS have largely been unfruitful due to the multi-valuedness nature (every shipment contains multiple identical pallets with different DoS) and data sparsity induced by lack of matching historical conditions. In this paper, we introduce a new framework for storage assignment that provides a solution to the DoS prediction problem through a distributional reformulation and a novel neural network, ParallelNet. Through collaboration with a world-leading cold storage company, we show that the system is able to predict DoS with a MAPE of 29%, a decrease of ~30% compared to a CNN-LSTM model, and suffers less performance decay into the future. The framework is then integrated into a first-of-its-kind Storage Assignment system, which is being deployed in warehouses across United States, with initial results showing up to 21% in labor savings. We also release the first publicly available set of warehousing records to facilitate research into this central problem.

****

CAQL: Continuous Action Q-Learning

Moonkyung Ryu,Yinlam Chow,Ross Anderson,Christian Tjandraatmadja,Craig Boutilier

Reinforcement learning (RL) with value-based methods (e.g., Q-learning) has shown success in a variety of domains such as
games and recommender systems (RSs). When the action space is finite, these algorithms implicitly finds a policy by learning the optimal value function, which are often very efficient.

However, one major challenge of extending Q-learning to tackle continuous-action RL problems is that obtaining optimal Bellman backup requires solving a continuous action-maximization (max-Q) problem. While it is common to restrict the parameterization of the Q-function to be concave in actions to simplify the max-Q problem, such a restriction might lead to performance degradation. Alternatively, when the Q-function is parameterized with a generic feed-forward neural network (NN), the max-Q problem can be NP-hard. In this work, we propose the CAQL method which minimizes the Bellman residual using Q-learning with one of several plug-and-play action optimizers. In particular, leveraging the strides of optimization theories in deep NN, we show that max-Q problem can be solved optimally with mixed-integer programming (MIP)---when the Q-function has sufficient representation power, this MIP-based optimization induces better policies and is more robust than counterparts, e.g., CEM or GA, that approximate the max-Q solution. To speed up training of CAQL, we develop three techniques, namely (i) dynamic tolerance, (ii) dual filtering, and (iii) clustering.

To speed up inference of CAQL, we introduce the action function that concurrently learns the optimal policy.

To demonstrate the efficiency of CAQL we compare it with state-of-the-art RL algorithms on benchmark continuous control problems that have different degrees of action constraints and show that CAQL significantly outperforms policy-based methods in heavily constrained environments.

****

GRAPH ANALYSIS AND GRAPH POOLING IN THE SPATIAL DOMAIN

Mostafa Rahmani,Ping Li

The spatial convolution layer which is widely used in the Graph Neural Networks (GNNs) aggregates the feature vector of each node with the feature vectors of its neighboring nodes. The GNN is not aware of the locations of the nodes in the global structure of the graph and when the local structures corresponding to different nodes are similar to each other, the convolution layer maps all those nodes to similar or same feature vectors in the continuous feature space. Therefore, the GNN cannot distinguish two graphs if their difference is not in their local structures. In addition, when the nodes are not labeled/attributed the convolution layers can fail to distinguish even different local structures. In this paper, we propose an effective solution to address this problem of the GNNs. The proposed approach leverages a spatial representation of the graph which makes the neural network aware of the differences between the nodes and also their locations in the graph. The spatial representation which is equivalent to a point-cloud representation of the graph is obtained by a graph embedding method. Using the proposed approach, the local feature extractor of the GNN distinguishes similar local structures in different locations of the graph and the GNN infers the topological structure of the graph from the spatial distribution of the locally extracted feature vectors. Moreover, the spatial representation is utilized to simplify the graph down-sampling problem. A new graph pooling method is proposed and it is shown that the proposed pooling method achieves competitive or better results in comparison with the state-of-the-art methods.

****
Your classifier is secretly an energy based model and you should treat it like one
Will Grathwohl,Kuan-Chieh Wang,Joern-Henrik Jacobsen,David Duvenaud,Mohammad Norouzi,Kevin Swersky
We propose to reinterpret a standard discriminative classifier of p(y|x) as an energy based model for the joint distribution p(x, y). In this setting, the standard class probabilities can be easily computed as well as unnormalized values of p(x) and p(x|y). Within this framework, standard discriminative architectures may be used and the model can also be trained on unlabeled data. We demonstrate that energy based training of the joint distribution improves calibration, robustness, and out-of-distribution detection while also enabling our models to generate samples rivaling the quality of recent GAN approaches. We improve upon recently proposed techniques for scaling up the training of energy based models and present an approach which adds little overhead compared to standard classification training. Our approach is the first to achieve performance rivaling the state-of-the-art in both generative and discriminative learning within one hybrid model.
****
On the Linguistic Capacity of Real-time Counter Automata
William Merrill
While counter machines have received little attention in theoretical computer science since the 1960s, they have recently achieved a newfound relevance to the field of natural language processing (NLP). Recent work has suggested that some strong-performing recurrent neural networks utilize their memory as counters. Thus, one potential way to understand the sucess of these networks is to revisit the theory of counter computation. Therefore, we choose to study the abilities of real-time counter machines as formal grammars. We first show that several variants of the counter machine converge to express the same class of formal languages. We also prove that counter languages are closed under complement, union, intersection, and many other common set operations. Next, we show that counter machines cannot evaluate boolean expressions, even though they can weakly validate their syntax. This has implications for the interpretability and evaluation of neural network systems: successfully matching syntactic patterns does not guarantee that a counter-like model accurately represents underlying semantic structures. Finally, we consider the question of whether counter languages are semilinear. This work makes general contributions to the theory of formal languages that are of particular interest for the interpretability of recurrent neural networks.

****
Combining MixMatch and Active Learning for Better Accuracy with Fewer Labels
Shuang Song,David Berthelot,Afshin Rostamizadeh
We propose using active learning based techniques to further improve the state-of-the-art semi-supervised learning MixMatch algorithm. We provide a thorough empirical evaluation of several active-learning and baseline methods, which successfully demonstrate a significant improvement on the benchmark CIFAR-10, CIFAR-100, and SVHN datasets (as much as 1.5% in absolute accuracy).
We also provide an empirical analysis of the cost trade-off between incrementally gathering more labeled versus unlabeled data. This analysis can be used to measure the relative value of labeled/unlabeled data at different points of the learning curve, where we find that although the incremental value of labeled data can be as much as 20x that of unlabeled, it quickly diminishes to less than 3x once more than 2,000 labeled example are observed.
****
Adaptive Structural Fingerprints for Graph Attention Networks
Kai Zhang,Yaokang Zhu,Jun Wang,Jie Zhang
Graph attention network (GAT) is a promising framework to perform convolution and massage passing on graphs. Yet, how to fully exploit rich structural information in the attention mechanism remains a challenge. In the current version, GAT calculates attention scores mainly using node features and among one-hop neighbors, while increasing the attention range to higher-order neighbors can negatively affect its performance, reflecting the over-smoothing risk of GAT (or graph neural networks in general), and the ineffectiveness in exploiting graph structural details. In this paper, we propose an ``"adaptive structural fingerprint" (ADSF) model to fully exploit graph topological details in graph attention network. The key idea is to contextualize each node with a weighted, learnable receptive field encoding rich and diverse local graph structures. By doing this, structural interactions between the nodes can be inferred accurately, thus significantly improving subsequent attention layer as well as the convergence of learning. Furthermore, our model provides a useful platform for different subspaces of node features and various scales of graph structures to ``cross-talk'' with each other through the learning of multi-head attention, being particularly useful in handling complex real-world data. Empirical results demonstrate the power of our approach in exploiting rich structural information in GAT and in alleviating the intrinsic oversmoothing problem in graph neural networks.
****
Inductive Matrix Completion Based on Graph Neural Networks
Muhan Zhang,Yixin Chen
We propose an inductive matrix completion model without using side information. By factorizing the (rating) matrix into the product of low-dimensional latent embeddings of rows (users) and columns (items), a majority of existing matrix completion methods are transductive, since the learned embeddings cannot generalize to unseen rows/columns or to new matrices. To make matrix completion inductive, most previous works use content (side information), such as user's age or movie's genre, to make predictions. However, high-quality content is not always available, and can be hard to extract. Under the extreme setting where not any side information is available other than the matrix to complete, can we still learn an inductive matrix completion model? In this paper, we propose an Inductive Graph-based Matrix Completion (IGMC) model to address this problem. IGMC trains a graph neural network (GNN) based purely on 1-hop subgraphs around (user, item) pairs generated from the rating matrix and maps these subgraphs to their corresponding ratings. It achieves highly competitive performance with state-of-the-art transductive baselines. In addition, IGMC is inductive -- it can generalize to users/items unseen during the training (given that their interactions exist), and can even transfer to new tasks. Our transfer learning experiments show that a model trained out of the MovieLens dataset can be directly used to predict Douban movie ratings with surprisingly good performance. Our work demonstrates that: 1) it is possible to train inductive matrix completion models without using side information while achieving similar or better performances than state-of-the-art tra

nsductive methods; 2) local graph patterns around a (user, item) pair are effect ive predictors of the rating this user gives to the item; and 3) Long-range depe ndencies might not be necessary for modeling recommender systems.
****
Neural Operator Search
Wei Li,Shaogang Gong,Xiatian Zhu
Existing neural architecture search (NAS) methods explore a limited feature-tran sformation-only search space while ignoring other advanced feature operations su ch as feature self-calibration by attention and dynamic convolutions. This disab les the NAS algorithms to discover more advanced network architectures. We addre ss this limitation by additionally exploiting feature self-calibration operation s, resulting in a heterogeneous search space. To solve the challenges of operati on heterogeneity and significantly larger search space, we formulate a neural op erator search (NOS) method. NOS presents a novel heterogeneous residual block fo r integrating the heterogeneous operations in a unified structure, and an attent ion guided search strategy for facilitating the search process over a vast space . Extensive experiments show that NOS can search novel cell architectures with h ighly competitive performance on the CIFAR and ImageNet benchmarks.

****
Time2Vec: Learning a Vector Representation of Time
Seyed Mehran Kazemi,Rishab Goel,Sepehr Eghbali,Janahan Ramanan,Jaspreet Sahota,S anjay Thakur,Stella Wu,Cathal Smyth,Pascal Poupart,Marcus Brubaker
Time is an important feature in many applications involving events that occur sy nchronously and/or asynchronously. To effectively consume time information, rece nt studies have focused on designing new architectures. In this paper, we take a n orthogonal but complementary approach by providing a model-agnostic vector rep resentation for time, called Time2Vec, that can be easily imported into many exi sting and future architectures and improve their performances. We show on a rang e of models and problems that replacing the notion of time with its Time2Vec rep resentation improves the performance of the final model.
****
ReMixMatch: Semi-Supervised Learning with Distribution Matching and Augmentation Anchoring
David Berthelot,Nicholas Carlini,Ekin D. Cubuk,Alex Kurakin,Kihyuk Sohn,Han Zhan g,Colin Raffel
We improve the recently-proposed ``MixMatch semi-supervised learning algorithm b y introducing two new techniques: distribution alignment and augmentation anchor ing.
- Distribution alignment encourages the marginal distribution of predictions on unlabeled data to be close to the marginal distribution of ground-truth labels.
- Augmentation anchoring} feeds multiple strongly augmented versions of an input into the model and encourages each output to be close to the prediction for a w eakly-augmented version of the same input.
To produce strong augmentations, we propose a variant of AutoAugment which learn s the augmentation policy while the model is being trained.

Our new algorithm, dubbed ReMixMatch, is significantly more data-efficient than prior work, requiring between 5 times and 16 times less data to reach the same a ccuracy. For example, on CIFAR-10 with 250 labeled examples we reach 93.73% accu racy (compared to MixMatch's accuracy of 93.58% with 4000 examples) and a median accuracy of 84.92% with just four labels per class.

****
Conditional Learning of Fair Representations
Han Zhao,Amanda Coston,Tameem Adel,Geoffrey J. Gordon
We propose a novel algorithm for learning fair representations that can simultan eously mitigate two notions of disparity among different demographic subgroups i n the classification setting. Two key components underpinning the design of our algorithm are balanced error rate and conditional alignment of representations.

We show how these two components contribute to ensuring accuracy parity and equalized false-positive and false-negative rates across groups without impacting demographic parity. Furthermore, we also demonstrate both in theory and on two real-world experiments that the proposed algorithm leads to a better utility-fairness trade-off on balanced datasets compared with existing algorithms on learning fair representations for classification.

****
Mean-field Behaviour of Neural Tangent Kernel for Deep Neural Networks
Soufiane Hayou,Arnaud Doucet,Judith Rousseau
Recent work by Jacot et al. (2018) has showed that training a neural network of any kind with gradient descent in parameter space is equivalent to kernel gradient descent in function space with respect to the Neural Tangent Kernel (NTK). Lee et al. (2019) built on this result to show that the output of a neural network trained using full batch gradient descent can be approximated by a linear model for wide networks. In parallel, a recent line of studies (Schoenhols et al. (2017), Hayou et al. (2019)) suggested that a special initialization known as the Edge of Chaos leads to good performance. In this paper, we bridge the gap between this two concepts and show the impact of the initialization and the activation function on the NTK as the network depth becomes large. We provide experiments illustrating our theoretical results.
****
TabNet: Attentive Interpretable Tabular Learning
Sercan O. Arik,Tomas Pfister
We propose a novel high-performance interpretable deep tabular data learning network, TabNet. TabNet utilizes a sequential attention mechanism that softly selects features to reason from at each decision step and then aggregates the processed information to make a final prediction decision. By explicitly selecting sparse features, TabNet learns very efficiently as the model capacity at each decision step is fully utilized for the most relevant features, resulting in a high performance model. This sparsity also enables more interpretable decision making through the visualization of feature selection masks. We demonstrate that TabNet outperforms other neural network and decision tree variants on a wide range of tabular data learning datasets and yields interpretable feature attributions and insights into the global model behavior.
****
Adapt-to-Learn: Policy Transfer in Reinforcement Learning
Girish Joshi,Girish Chowdhary
Efficient and robust policy transfer remains a key challenge in reinforcement learning. Policy transfer through warm initialization, imitation, or interacting over a large set of agents with randomized instances, have been commonly applied to solve a variety of Reinforcement Learning (RL) tasks. However, this is far from how behavior transfer happens in the biological world: Humans and animals are able to quickly adapt the learned behaviors between similar tasks and learn new skills when presented with new situations. Here we seek to answer the question: Will learning to combine adaptation reward with environmental reward lead to a more efficient transfer of policies between domains? We introduce a principled mechanism that can \textbf{``Adapt-to-Learn"}, that is adapt the source policy to learn to solve a target task with significant transition differences and uncertainties. We show through theory and experiments that our method leads to a significantly reduced sample complexity of transferring the policies between the tasks.
****
Identity Crisis: Memorization and Generalization Under Extreme Overparameterization
Chiyuan Zhang,Samy Bengio,Moritz Hardt,Michael C. Mozer,Yoram Singer
We study the interplay between memorization and generalization of overparameterized networks in the extreme case of a single training example and an identity-mapping task. We examine fully-connected and convolutional networks (FCN and CNN), both linear and nonlinear, initialized randomly and then trained

to minimize the reconstruction error. The trained networks stereotypically take one of two forms: the constant function (memorization) and the identity function (generalization).
We formally characterize generalization in single-layer FCNs and CNNs.
We show empirically that different architectures exhibit strikingly different inductive biases.
For example, CNNs of up to 10 layers are able to generalize
from a single example, whereas FCNs cannot learn the identity function reliably from 60k examples. Deeper CNNs often fail, but nonetheless do astonishing work to memorize the training output: because CNN biases are location invariant, the model must progressively grow an output pattern from the image boundaries via the coordination of many layers. Our work helps to quantify and visualize the sensitivity of inductive biases to architectural choices such as depth, kernel width, and number of channels.

****
Stiffness: A New Perspective on Generalization in Neural Networks
Stanislav Fort,Pawe■ Krzysztof Nowak,Stanis■aw Jastrzebski,Srini Narayanan
We investigate neural network training and generalization using the concept of stiffness. We measure how stiff a network is by looking at how a small gradient step on one example affects the loss on another example. In particular, we study how stiffness depends on 1) class membership, 2) distance between data points in the input space, 3) training iteration, and 4) learning rate. We experiment on MNIST, FASHION MNIST, and CIFAR-10 using fully-connected and convolutional neural networks. Our results demonstrate that stiffness is a useful concept for diagnosing and characterizing generalization. We observe that small learning rates reliably lead to higher stiffness at a given epoch as well as at a given training loss. In addition, we measure how stiffness between two data points depends on their mutual input-space distance, and establish the concept of a dynamical critical length that characterizes the distance over which datapoints react similarly to gradient updates. The dynamical critical length decreases with training and the higher the learning rate, the smaller the critical length.
****
Linguistic Embeddings as a Common-Sense Knowledge Repository: Challenges and Opportunities
Nancy Fulda
Many applications of linguistic embedding models rely on their value as pre-trained inputs for end-to-end tasks such as dialog modeling, machine translation, or question answering. This position paper presents an alternate paradigm: Rather than using learned embeddings as input features, we instead treat them as a common-sense knowledge repository that can be queried via simple mathematical operations within the embedding space. We show how linear offsets can be used to (a) identify an object given its description, (b) discover relations of an object given its label, and (c) map free-form text to a set of action primitives. Our experiments provide a valuable proof of concept that language-informed common sense reasoning, or `reasoning in the linguistic domain', lies within the grasp of the research community. In order to attain this goal, however, we must reconsider the way neural embedding models are typically trained an evaluated. To that end, we also identify three empirically-motivated evaluation metrics for use in the training of future embedding models.
****
First-Order Preconditioning via Hypergradient Descent
Ted Moskovitz,Rui Wang,Janice Lan,Sanyam Kapoor,Thomas Miconi,Jason Yosinski,Aditya Rawal
Standard gradient-descent methods are susceptible to a range of issues that can impede training, such as high correlations and different scaling in parameter space. These difficulties can be addressed by second-order approaches that apply a preconditioning matrix to the gradient to improve convergence. Unfortunately, such algorithms typically struggle to scale to high-dimensional problems, in part because the calculation of specific preconditioners such as the inverse Hessia

n or Fisher information matrix is highly expensive. We introduce first-order pre conditioning (FOP), a fast, scalable approach that generalizes previous work on hypergradient descent (Almeida et al., 1998; Maclaurin et al., 2015; Baydin et a l., 2017) to learn a preconditioning matrix that only makes use of first-order i nformation. Experiments show that FOP is able to improve the performance of stan dard deep learning optimizers on several visual classification tasks with minima l computational overhead. We also investigate the properties of the learned prec onditioning matrices and perform a preliminary theoretical analysis of the algor ithm.
****

Feature Partitioning for Efficient Multi-Task Architectures
Alejandro Newell,Lu Jiang,Chong Wang,Li-Jia Li,Jia Deng
Multi-task learning promises to use less data, parameters, and time than trainin g separate single-task models. But realizing these benefits in practice is chall enging. In particular, it is difficult to define a suitable architecture that ha s enough capacity to support many tasks while not requiring excessive compute fo r each individual task. There are difficult trade-offs when deciding how to allo cate parameters and layers across a large set of tasks. To address this, we prop ose a method for automatically searching over multi-task architectures that acco unts for resource constraints. We define a parameterization of feature sharing s trategies for effective coverage and sampling of architectures. We also present a method for quick evaluation of such architectures with feature distillation. T ogether these contributions allow us to quickly optimize for parameter-efficient multi-task models. We benchmark on Visual Decathlon, demonstrating that we can automatically search for and identify architectures that effectively make trade- offs between task resource requirements while maintaining a high level of final performance.
****

Layer Flexible Adaptive Computation Time for Recurrent Neural Networks
Lida Zhang,Diego Klabjan
Deep recurrent neural networks perform well on sequence data and are the model o f choice. However, it is a daunting task to decide the structure of the networks , i.e. the number of layers, especially considering different computational need s of a sequence. We propose a layer flexible recurrent neural network with adapt ive computation time, and expand it to a sequence to sequence model. Different f rom the adaptive computation time model, our model has a dynamic number of trans mission states which vary by step and sequence. We evaluate the model on a finan cial data set and Wikipedia language modeling. Experimental results show the per formance improvement of 7% to 12% and indicate the model's ability to dynamicall y change the number of layers along with the computational steps.
****

Curvature-based Robustness Certificates against Adversarial Examples
Sahil Singla,Soheil Feizi
A robustness certificate against adversarial examples is the minimum distance of a given input to the decision boundary of the classifier (or its lower bound). For {\it any} perturbation of the input with a magnitude smaller than the certif icate value, the classification output will provably remain unchanged. Computing exact robustness certificates for deep classifiers is difficult in general sinc e it requires solving a non-convex optimization. In this paper, we provide compu tationally-efficient robustness certificates for deep classifiers with different iable activation functions in two steps. First, we show that if the eigenvalues of the Hessian of the network (curvatures of the network) are bounded, we can co mpute a robustness certificate in the l_2 norm efficiently using convex optimi zation. Second, we derive a computationally-efficient differentiable upper bound on the curvature of a deep network. We also use the curvature bound as a regula rization term during the training of the network to boost its certified robustne ss against adversarial examples. Putting these results together leads to our pro posed {\bf C}urvature-based {\bf R}obustness {\bf C}ertificate (CRC) and {\bf C} urvature-based {\bf R}obust {\bf T}raining (CRT). Our numerical results show tha t CRC outperforms CROWN's certificate by an order of magnitude while CRT leads t

o higher certified accuracy compared to standard adversarial training and TRADES
.

****
Adversarial Video Generation on Complex Datasets
Aidan Clark,Jeff Donahue,Karen Simonyan
Generative models of natural images have progressed towards high fidelity sample
s by the strong leveraging of scale. We attempt to carry this success to the fie
ld of video modeling by showing that large Generative Adversarial Networks train
ed on the complex Kinetics-600 dataset are able to produce video samples of subs
tantially higher complexity and fidelity than previous work. Our proposed model
, Dual Video Discriminator GAN (DVD-GAN), scales to longer and higher resolution
 videos by leveraging a computationally efficient decomposition of its discrimin
ator. We evaluate on the related tasks of video synthesis and video prediction,
 and achieve new state-of-the-art Fréchet Inception Distance for prediction for
Kinetics-600, as well as state-of-the-art Inception Score for synthesis on the
UCF-101 dataset, alongside establishing a strong baseline for synthesis on Kinet
ics-600.
****
Topological Autoencoders
Michael Moor,Max Horn,Bastian Rieck,Karsten Borgwardt
We propose a novel approach for preserving topological structures of the input s
pace in latent representations of autoencoders. Using persistent homology, a tec
hnique from topological data analysis, we calculate topological signatures of bo
th the input and latent space to derive a topological loss term. Under weak theo
retical assumptions, we can construct this loss in a differentiable manner, such
 that the encoding learns to retain multi-scale connectivity information.
We show that our approach is theoretically well-founded and that it exhibits fav
ourable latent representations on a synthetic manifold as well as on real-world
image data sets, while preserving low reconstruction errors.
****
Reinforcement Learning without Ground-Truth State
Xingyu Lin,Harjatin Singh Baweja,David Held
To perform robot manipulation tasks, a low-dimensional state of the environment
typically needs to be estimated. However, designing a state estimator can someti
mes be difficult, especially in environments with deformable objects. An alterna
tive is to learn an end-to-end policy that maps directly from high-dimensional s
ensor inputs to actions. However, if this policy is trained with reinforcement l
earning, then without a state estimator, it is hard to specify a reward function
 based on high-dimensional observations. To meet this challenge, we propose a si
mple indicator reward function for goal-conditioned reinforcement learning: we o
nly give a positive reward when the robot's observation exactly matches a target
 goal observation. We show that by relabeling the original goal with the achieve
d goal to obtain positive rewards (Andrychowicz et al., 2017), we can learn with
 the indicator reward function even in continuous state spaces. We propose two m
ethods to further speed up convergence with indicator rewards: reward balancing
and reward filtering. We show comparable performance between our method and an o
racle which uses the ground-truth state for computing rewards. We show that our
method can perform complex tasks in continuous state spaces such as rope manipul
ation from RGB-D images, without knowledge of the ground-truth state.
****
Improved Sample Complexities for Deep Neural Networks and Robust Classification
via an All-Layer Margin
Colin Wei,Tengyu Ma
For linear classifiers, the relationship between (normalized) output margin and
generalization is captured in a clear and simple bound – a large output margin i
mplies good generalization. Unfortunately, for deep models, this relationship is
 less clear: existing analyses of the output margin give complicated bounds whic
h sometimes depend exponentially on depth. In this work, we propose to instead a
nalyze a new notion of margin, which we call the “all-layer margin.” Our analysi

s reveals that the all-layer margin has a clear and direct relationship with gen
eralization for deep models. This enables the following concrete applications of
 the all-layer margin: 1) by analyzing the all-layer margin, we obtain tighter g
eneralization bounds for neural nets which depend on Jacobian and hidden layer n
orms and remove the exponential dependency on depth 2) our neural net results ea
sily translate to the adversarially robust setting, giving the first direct anal
ysis of robust test error for deep networks, and 3) we present a theoretically i
nspired training algorithm for increasing the all-layer margin. Our algorithm im
proves both clean and adversarially robust test performance over strong baseline
s in practice.
****

In-Domain Representation Learning For Remote Sensing
Maxim Neumann,Andre Susano Pinto,Xiaohua Zhai,Neil Houlsby
Given the importance of remote sensing, surprisingly little attention has been p
aid to it by the representation learning community. To address it and to speed u
p innovation in this domain, we provide simplified access to 5 diverse remote se
nsing datasets in a standardized form. We specifically explore in-domain represe
ntation learning and address the question of "what characteristics should a data
set have to be a good source for remote sensing representation learning". The es
tablished baselines achieve state-of-the-art performance on these datasets.

****

Training Neural Networks for and by Interpolation
Leonard Berrada,Andrew Zisserman,Pawan M. Kumar
In modern supervised learning, many deep neural networks are able to interpolate
 the data: the empirical loss can be driven to near zero on all samples simultan
eously. In this work, we explicitly exploit this interpolation property for the
design of a new optimization algorithm for deep learning. Specifically, we use i
t to compute an adaptive learning-rate in closed form at each iteration. This re
sults in the Adaptive Learning-rates for Interpolation with Gradients (ALI-G) al
gorithm. ALI-G retains the main advantage of SGD which is a low computational co
st per iteration. But unlike SGD, the learning-rate of ALI-G uses a single const
ant hyper-parameter and does not require a decay schedule, which makes it consid
erably easier to tune. We provide convergence guarantees of ALI-G in the stochas
tic convex setting. Notably, all our convergence results tackle the realistic ca
se where the interpolation property is satisfied up to some tolerance. We provid
e experiments on a variety of architectures and tasks: (i) learning a differenti
able neural computer; (ii) training a wide residual network on the SVHN data set
; (iii) training a Bi-LSTM on the SNLI data set; and (iv) training wide residual
 networks and densely connected networks on the CIFAR data sets. ALI-G produces
state-of-the-art results among adaptive methods, and even yields comparable perf
ormance with SGD, which requires manually tuned learning-rate schedules. Further
more, ALI-G is simple to implement in any standard deep learning framework and c
an be used as a drop-in replacement in existing code.
****

Unsupervised Data Augmentation for Consistency Training
Qizhe Xie,Zihang Dai,Eduard Hovy,Minh-Thang Luong,Quoc V. Le
Semi-supervised learning lately has shown much promise in improving deep learnin
g models when labeled data is scarce. Common among recent approaches is the use
of consistency training on a large amount of unlabeled data to constrain model p
redictions to be invariant to input noise. In this work, we present a new perspe
ctive on how to effectively noise unlabeled examples and argue that the quality
of noising, specifically those produced by advanced data augmentation methods, p
lays a crucial role in semi-supervised learning. By substituting simple noising
operations with advanced data augmentation methods, our method brings substantia
l improvements across six language and three vision tasks under the same consist
ency training framework. On the IMDb text classification dataset, with only 20 l
abeled examples, our method achieves an error rate of 4.20, outperforming the st
ate-of-the-art model trained on 25,000 labeled examples. On a standard semi-supe
rvised learning benchmark, CIFAR-10, our method outperforms all previous approac

hes and achieves an error rate of 2.7% with only 4,000 examples, nearly matching the performance of models trained on 50,000 labeled examples. Our method also combines well with transfer learning, e.g., when finetuning from BERT, and yields improvements in high-data regime, such as ImageNet, whether when there is only 10% labeled data or when a full labeled set with 1.3M extra unlabeled examples is used.
****

Assessing Generalization in TD methods for Deep Reinforcement Learning
Emmanuel Bengio,Doina Precup,Joelle Pineau
Current Deep Reinforcement Learning (DRL) methods can exhibit both data inefficiency and brittleness, which seem to indicate that they generalize poorly. In this work, we experimentally analyze this issue through the lens of memorization, and show that it can be observed directly during training. More precisely, we find that Deep Neural Networks (DNNs) trained with supervised tasks on trajectories capture temporal structure well, but DNNs trained with TD(0) methods struggle to do so, while using TD(lambda) targets leads to better generalization.
****

Cyclical Stochastic Gradient MCMC for Bayesian Deep Learning
Ruqi Zhang,Chunyuan Li,Jianyi Zhang,Changyou Chen,Andrew Gordon Wilson
The posteriors over neural network weights are high dimensional and multimodal. Each mode typically characterizes a meaningfully different representation of the data. We develop Cyclical Stochastic Gradient MCMC (SG-MCMC) to automatically explore such distributions. In particular, we propose a cyclical stepsize schedule, where larger steps discover new modes, and smaller steps characterize each mode. We prove non-asymptotic convergence theory of our proposed algorithm. Moreover, we provide extensive experimental results, including ImageNet, to demonstrate the effectiveness of cyclical SG-MCMC in learning complex multimodal distributions, especially for fully Bayesian inference with modern deep neural networks.
****

Why Does Hierarchy (Sometimes) Work So Well in Reinforcement Learning?
Ofir Nachum,Haoran Tang,Xingyu Lu,Shixiang Gu,Honglak Lee,Sergey Levine
Hierarchical reinforcement learning has demonstrated significant success at solving difficult reinforcement learning (RL) tasks. Previous works have motivated the use of hierarchy by appealing to a number of intuitive benefits, including learning over temporally extended transitions, exploring over temporally extended periods, and training and exploring in a more semantically meaningful action space, among others. However, in fully observed, Markovian settings, it is not immediately clear why hierarchical RL should provide benefits over standard "shallow" RL architectures. In this work, we isolate and evaluate the claimed benefits of hierarchical RL on a suite of tasks encompassing locomotion, navigation, and manipulation.
Surprisingly, we find that most of the observed benefits of hierarchy can be attributed to improved exploration, as opposed to easier policy learning or imposed hierarchical structures. Given this insight, we present exploration techniques inspired by hierarchy that achieve performance competitive with hierarchical RL while at the same time being much simpler to use and implement.
****

The Effect of Neural Net Architecture on Gradient Confusion & Training Performance
Karthik A. Sankararaman,Soham De,Zheng Xu,W. Ronny Huang,Tom Goldstein
The goal of this paper is to study why typical neural networks train so fast, and how neural network architecture affects the speed of training. We introduce a simple concept called gradient confusion to help formally analyze this. When confusion is high, stochastic gradients produced by different data samples may be negatively correlated, slowing down convergence. But when gradient confusion is low, data samples interact harmoniously, and training proceeds quickly. Through novel theoretical and experimental results, we show how the neural net architecture affects gradient confusion, and thus the efficiency of training. We show that increasing the width of neural networks leads to lower gradient confusion, and thus easier model training. On the other hand, increasing the depth of neural n

etworks has the opposite effect. Finally, we observe empirically that techniques like batch normalization and skip connections reduce gradient confusion, which helps reduce the training burden of very deep networks.
****

Augmenting Genetic Algorithms with Deep Neural Networks for Exploring the Chemical Space
AkshatKumar Nigam,Pascal Friederich,Mario Krenn,Alan Aspuru-Guzik
Challenges in natural sciences can often be phrased as optimization problems. Machine learning techniques have recently been applied to solve such problems. One example in chemistry is the design of tailor-made organic materials and molecules, which requires efficient methods to explore the chemical space. We present a genetic algorithm (GA) that is enhanced with a neural network (DNN) based discriminator model to improve the diversity of generated molecules and at the same time steer the GA. We show that our algorithm outperforms other generative models in optimization tasks. We furthermore present a way to increase interpretability of genetic algorithms, which helped us to derive design principles
****

Regularizing Deep Multi-Task Networks using Orthogonal Gradients
Mihai Suteu,Yi-ke Guo
Deep neural networks are a promising approach towards multi-task learning because of their capability to leverage knowledge across domains and learn general purpose representations. Nevertheless, they can fail to live up to these promises as tasks often compete for a model's limited resources, potentially leading to lower overall performance. In this work we tackle the issue of interfering tasks through a comprehensive analysis of their training, derived from looking at the interaction between gradients within their shared parameters. Our empirical results show that well-performing models have low variance in the angles between task gradients and that popular regularization methods implicitly reduce this measure. Based on this observation, we propose a novel gradient regularization term that minimizes task interference by enforcing near orthogonal gradients. Updating the shared parameters using this property encourages task specific decoders to optimize different parts of the feature extractor, thus reducing competition. We evaluate our method with classification and regression tasks on the multiDigitMNIST and NYUv2 dataset where we obtain competitive results. This work is a first step towards non-interfering multi-task optimization.
****

Fast Training of Sparse Graph Neural Networks on Dense Hardware
Matej Balog,Bart van Merriënboer,Subhodeep Moitra,Yujia Li,Daniel Tarlow
Graph neural networks have become increasingly popular in recent years due to their ability to naturally encode relational input data and their ability to operate on large graphs by using a sparse representation of graph adjacency matrices. As we look to scale up these models using custom hardware, a natural assumption would be that we need hardware tailored to sparse operations and/or dynamic control flow. In this work, we question this assumption by scaling up sparse graph neural networks using a platform targeted at dense computation on fixed-size data. Drawing inspiration from optimization of numerical algorithms on sparse matrices, we develop techniques that enable training the sparse graph neural network model from Allamanis et al. (2018) in 13 minutes using a 512-core TPUv2 Pod, whereas the original training takes almost a day.
****

Simultaneous Classification and Out-of-Distribution Detection Using Deep Neural Networks
Aristotelis-Angelos Papadopoulos,Nazim Shaikh,Jiamian Wang,Mohammad Reza Rajati
Deep neural networks have achieved great success in classi■cation tasks during the last years. However, one major problem to the path towards arti■cial intelligence is the inability of neural networks to accurately detect samples from novel class distributions and therefore, most of the existent classi■cation algorithms assume that all classes are known prior to the training stage. In this work, we propose a methodology for training a neural network that allows it to ef■ciently detect out-of-distribution (OOD) examples without compromising much of its cl

assi■cation accuracy on the test examples from known classes. Based on the Outli
er Exposure (OE) technique, we propose a novel loss function that achieves state
-of-the-art results in out-of-distribution detection with OE both on image and t
ext classi■cation tasks. Additionally, the way this method was constructed makes
 it suitable for training any classi■cation algorithm that is based on Maximum L
ikelihood methods.
****

Rapid Learning or Feature Reuse? Towards Understanding the Effectiveness of MAML
Aniruddh Raghu,Maithra Raghu,Samy Bengio,Oriol Vinyals
An important research direction in machine learning has centered around developi
ng meta-learning algorithms to tackle few-shot learning. An especially successfu
l algorithm has been Model Agnostic Meta-Learning (MAML), a method that consists
 of two optimization loops, with the outer loop finding a meta-initialization, f
rom which the inner loop can efficiently learn new tasks. Despite MAML's popular
ity, a fundamental open question remains -- is the effectiveness of MAML due to
the meta-initialization being primed for rapid learning (large, efficient change
s in the representations) or due to feature reuse, with the meta initialization
 already containing high quality features? We investigate this question, via abl
ation studies and analysis of the latent representations, finding that feature r
euse is the dominant factor. This leads to the ANIL (Almost No Inner Loop) algor
ithm, a simplification of MAML where we remove the inner loop for all but the (t
ask-specific) head of the underlying neural network. ANIL matches MAML's perform
ance on benchmark few-shot image classification and RL and offers computational
improvements over MAML. We further study the precise contributions of the head a
nd body of the network, showing that performance on the test tasks is entirely d
etermined by the quality of the learned features, and we can remove even the hea
d of the network (the NIL algorithm). We conclude with a discussion of the rapid
 learning vs feature reuse question for meta-learning algorithms more broadly.
****

Long-term planning, short-term adjustments
Hamed Khorasgani,Chi Zhang,Chetan Gupta,Susumu Serita
Deep Reinforcement Learning (RL) algorithms can learn complex policies to optimi
ze
agent operation over time. RL algorithms have shown promising results
in solving complicated problems in recent years. However, their application on
real-world physical systems remains limited. Despite the advancements in RL
algorithms, the industries often prefer traditional control strategies. Traditio
nal
methods are simple, computationally efficient and easy to adjust. In this paper,
we propose a new Q-learning algorithm for continuous action space, which can
bridge the control and RL algorithms and bring us the best of both worlds. Our
method can learn complex policies to achieve long-term goals and at the same tim
e
it can be easily adjusted to address short-term requirements without retraining.
We achieve this by modeling both short-term and long-term prediction models.
The short-term prediction model represents the estimation of the system dynamic
while the long-term prediction model represents the Q-value. The case studies
demonstrate that our proposed method can achieve short-term and long-term goals
without complex reward functions.
****

Imitation Learning via Off-Policy Distribution Matching
Ilya Kostrikov,Ofir Nachum,Jonathan Tompson
When performing imitation learning from expert demonstrations, distribution matc
hing is a popular approach, in which one alternates between estimating distribut
ion ratios and then using these ratios as rewards in a standard reinforcement le
arning (RL) algorithm. Traditionally, estimation of the distribution ratio requi
res on-policy data, which has caused previous work to either be exorbitantly dat
a- inefficient or alter the original objective in a manner that can drastically
change its optimum. In this work, we show how the original distribution ratio es
timation objective may be transformed in a principled manner to yield a complete

ly off-policy objective. In addition to the data-efficiency that this provides, we are able to show that this objective also renders the use of a separate RL optimization unnecessary. Rather, an imitation policy may be learned directly from this objective without the use of explicit rewards. We call the resulting algorithm ValueDICE and evaluate it on a suite of popular imitation learning benchmarks, finding that it can achieve state-of-the-art sample efficiency and performance.
****

Unsupervised Learning of Automotive 3D Crash Simulations using LSTMs
Amin Abbasloo,Jochen Garcke,Rodrigo Iza-Teran
Long short-term memory (LSTM) networks allow to exhibit temporal dynamic behavior with feedback connections and seem a natural choice for learning sequences of 3D meshes. We introduce an approach for dynamic mesh representations as used for numerical simulations of car crashes. To bypass the complication of using 3D meshes, we transform the surface mesh sequences into spectral descriptors that efficiently encode the shape. A two branch LSTM based network architecture is chosen to learn the representations and dynamics of the crash during the simulation. The architecture is based on unsupervised video prediction by an LSTM without any convolutional layer. It uses an encoder LSTM to map an input sequence into a fixed length vector representation. On this representation one decoder LSTM performs the reconstruction of the input sequence, while the other decoder LSTM predicts the future behavior by receiving initial steps of the sequence as seed. The spatio-temporal error behavior of the model is analysed to study how well the model can extrapolate the learned spectral descriptors into the future, that is, how well it has learned to represent the underlying dynamical structural mechanics. Considering that only a few training examples are available, which is the typical case for numerical simulations, the network performs very well.
****

Augmenting Transformers with KNN-Based Composite Memory
Angela Fan,Claire Gardent,Chloe Braud,Antoine Bordes
Various machine learning tasks can benefit from access to external information of different modalities, such as text and images. Recent work has focused on learning architectures with large memories capable of storing this knowledge. We propose augmenting Transformer neural networks with KNN-based Information Fetching (KIF) modules. Each KIF module learns a read operation to access fixed external knowledge. We apply these modules to generative dialogue modeling, a challenging task where information must be flexibly retrieved and incorporated to maintain the topic and flow of conversation. We demonstrate the effectiveness of our approach by identifying relevant knowledge from Wikipedia, images, and human-written dialogue utterances, and show that leveraging this retrieved information improves model performance, measured by automatic and human evaluation.
****

SGD with Hardness Weighted Sampling for Distributionally Robust Deep Learning
Lucas Fidon,Sebastien Ourselin,Tom Vercauteren
Distributionally Robust Optimization (DRO) has been proposed as an alternative to Empirical Risk Minimization (ERM) in order to account for potential biases in the training data distribution. However, its use in deep learning has been severely restricted due to the relative inefficiency of the optimizers available for DRO compared to the wide-spread Stochastic Gradient Descent (SGD) based optimizers for deep learning with ERM. In this work, we demonstrate that SGD with hardness weighted sampling is a principled and efficient optimization method for DRO in machine learning and is particularly suited in the context of deep learning. Similar to a hard example mining strategy in essence and in practice, the proposed algorithm is straightforward to implement and computationally as efficient as SGD-based optimizers used for deep learning. It only requires adding a softmax layer and maintaining an history of the loss values for each training example to compute adaptive sampling probabilities. In contrast to typical ad hoc hard mining approaches, and exploiting recent theoretical results in deep learning optimization, we prove the convergence of our DRO algorithm for over-parameterized deep learning networks with ReLU activation and finite number of layers and p

arameters. Preliminary results demonstrate the feasibility and usefulness of ou
r approach.
****
Constrained Markov Decision Processes via Backward Value Functions
Harsh Satija,Philip Amortila,Joelle Pineau
Although Reinforcement Learning (RL) algorithms have found tremendous success in
 simulated domains, they often cannot directly be applied to physical systems, e
specially in cases where there are hard constraints to satisfy (e.g. on safety o
r resources). In standard RL, the agent is incentivized to explore any behavior
as long as it maximizes rewards, but in the real world undesired behavior can da
mage either the system or the agent in a way that breaks the learning process it
self. In this work, we model the problem of learning with constraints as a Const
rained Markov Decision Process, and provide a new on-policy formulation for solv
ing it. A key contribution of our approach is to translate cumulative cost const
raints into state-based constraints. Through this, we define a safe policy impro
vement method which maximizes returns while ensuring that the constraints are sa
tisfied at every step. We provide theoretical guarantees under which the agent c
onverges while ensuring safety over the course of training. We also highlight co
mputational advantages of this approach. The effectiveness of our approach is de
monstrated on safe navigation tasks and in safety-constrained versions of MuJoCo
 environments, with deep neural networks.
****
Reanalysis of Variance Reduced Temporal Difference Learning
Tengyu Xu,Zhe Wang,Yi Zhou,Yingbin Liang
Temporal difference (TD) learning is a popular algorithm for policy evaluation i
n reinforcement learning, but the vanilla TD can substantially suffer from the i
nherent optimization variance. A variance reduced TD (VRTD) algorithm was propos
ed by \cite{korda2015td}, which applies the variance reduction technique directl
y to the online TD learning with Markovian samples. In this work, we first point
 out the technical errors in the analysis of VRTD in \cite{korda2015td}, and the
n provide a mathematically solid analysis of the non-asymptotic convergence of V
RTD and its variance reduction performance. We show that VRTD is guaranteed to c
onverge to a neighborhood of the fixed-point solution of TD at a linear converge
nce rate. Furthermore, the variance error (for both i.i.d.\ and Markovian sampli
ng) and the bias error (for Markovian sampling) of VRTD are significantly reduce
d by the batch size of variance reduction in comparison to those of vanilla TD.
As a result, the overall computational complexity of VRTD to attain a given accu
rate solution outperforms that of TD under Markov sampling and outperforms that
of TD under i.i.d.\ sampling for a sufficiently small conditional number.
****
Meta-Learning for Variational Inference
Ruqi Zhang,Yingzhen Li,Chris De Sa,Sam Devlin,Cheng Zhang
Variational inference (VI) plays an essential role in approximate Bayesian infer
ence due to its computational efficiency and general applicability.
Crucial to the performance of VI is the selection of the divergence measure in t
he optimization objective, as it affects the properties of the approximate poste
rior significantly. In this paper, we propose a meta-learning algorithm to learn
 (i) the divergence measure suited for the task of interest to automate the desi
gn of the VI method; and (ii) initialization of the variational parameters, whic
h reduces the number of VI optimization steps drastically. We demonstrate the le
arned divergence outperforms the hand-designed divergence on Gaussian mixture di
stribution approximation, Bayesian neural network regression, and partial variat
ional autoencoder based recommender systems.
****
CONFEDERATED MACHINE LEARNING ON HORIZONTALLY AND VERTICALLY SEPARATED MEDICAL D
ATA FOR LARGE-SCALE HEALTH SYSTEM INTELLIGENCE
Dianbo Liu,Tim Miller,Kenneth Mandl
A patient’s health information is generally fragmented across silos. Though it i
s technically feasible to unite data for analysis in a manner that underpins a r
apid learning healthcare system, privacy concerns and regulatory barriers limit

data centralization. Machine learning can be conducted in a federated manner on patient datasets with the same set of variables, but separated across sites of care. But federated learning cannot handle the situation where different data types for a given
patient are separated vertically across different organizations. We call methods that enable machine learning model training on data separated by two or more degrees “confederated machine learning.” We built and evaluated a confederated machine
learning model to stratify the risk of accidental falls among the elderly.
****

Defending Against Adversarial Examples by Regularized Deep Embedding
Yao Li,Martin Renqiang Min,Wenchao Yu,Cho-Jui Hsieh,Thomas Lee,Erik Kruus
Recent studies have demonstrated the vulnerability of deep convolutional neural networks against adversarial examples. Inspired by the observation that the intrinsic dimension of image data is much smaller than its pixel space dimension and the vulnerability of neural networks grows with the input dimension, we propose to embed high-dimensional input images into a low-dimensional space to perform classification. However, arbitrarily projecting the input images to a low-dimensional space without regularization will not improve the robustness of deep neural networks. We propose a new framework, Embedding Regularized Classifier (ER-Classifier), which improves the adversarial robustness of the classifier through embedding regularization. Experimental results on several benchmark datasets show that, our proposed framework achieves state-of-the-art performance against strong adversarial attack methods.
****

Minimizing FLOPs to Learn Efficient Sparse Representations
Biswajit Paria,Chih-Kuan Yeh,Ian E.H. Yen,Ning Xu,Pradeep Ravikumar,Barnabás Póczos
Deep representation learning has become one of the most widely adopted approaches for visual search, recommendation, and identification. Retrieval of such representations from a large database is however computationally challenging. Approximate methods based on learning compact representations, have been widely explored for this problem, such as locality sensitive hashing, product quantization, and PCA. In this work, in contrast to learning compact representations, we propose to learn high dimensional and sparse representations that have similar representational capacity as dense embeddings while being more efficient due to sparse matrix multiplication operations which can be much faster than dense multiplication. Following the key insight that the number of operations decreases quadratically with the sparsity of embeddings provided the non-zero entries are distributed uniformly across dimensions, we propose a novel approach to learn such distributed sparse embeddings via the use of a carefully constructed regularization function that directly minimizes a continuous relaxation of the number of floating-point operations (FLOPs) incurred during retrieval. Our experiments show that our approach is competitive to the other baselines and yields a similar or better speed-vs-accuracy tradeoff on practical datasets.
****

Neural-Guided Symbolic Regression with Asymptotic Constraints
Li Li,Minjie Fan,Rishabh Singh,Patrick Riley
Symbolic regression is a type of discrete optimization problem that involves searching expressions that fit given data points. In many cases, other mathematical constraints about the unknown expression not only provide more information beyond just values at some inputs, but also effectively constrain the search space. We identify the asymptotic constraints of leading polynomial powers as the function approaches 0 and infinity as useful constraints and create a system to use them for symbolic regression. The first part of the system is a conditional expression generating neural network which preferentially generates expressions with the desired leading powers, producing novel expressions outside the training domain. The second part, which we call Neural-Guided Monte Carlo Tree Search, uses the network during a search to find an expression that conforms to a set of data points and desired leading powers. Lastly, we provide an extensive experimental

validation on thousands of target expressions showing the efficacy of our system compared to exiting methods for finding unknown functions outside of the training set.
****
Policy Optimization In the Face of Uncertainty
Tung-Long Vuong,Han Nguyen,Hai Pham,Kenneth Tran
Model-based reinforcement learning has the potential to be more sample efficient than model-free approaches. However, existing model-based methods are vulnerable to model bias, which leads to poor generalization and asymptotic performance compared to model-free counterparts. In this paper, we propose a novel policy optimization framework using an uncertainty-aware objective function to handle those issues. In this framework, the agent simultaneously learns an uncertainty-aware dynamics model and optimizes the policy according to these learned models. Under this framework, the objective function can represented end-to-end as a single computational graph, which allows seamless policy gradient computation via back propagation through the models. In addition to being theoretically sound, our approach shows promising results on challenging continuous control benchmarks with competitive asymptotic performance and sample complexity compared to state-of-the-art baselines.
****
Understanding Top-k Sparsification in Distributed Deep Learning
Shaohuai Shi,Xiaowen Chu,Ka Chun Cheung,Simon See
Distributed stochastic gradient descent (SGD) algorithms are widely deployed in training large-scale deep learning models, while the communication overhead among workers becomes the new system bottleneck. Recently proposed gradient sparsification techniques, especially Top-k sparsification with error compensation (TopK-SGD), can significantly reduce the communication traffic without obvious impact on the model accuracy. Some theoretical studies have been carried out to analyze the convergence property of TopK-SGD. However, existing studies do not dive into the details of Top-k operator in gradient sparsification and use relaxed bounds (e.g., exact bound of Random-k) for analysis; hence the derived results cannot well describe the real convergence performance of TopK-SGD. To this end, we first study the gradient distributions of TopK-SGD during training process through extensive experiments. We then theoretically derive a tighter bound for the Top-k operator. Finally, we exploit the property of gradient distribution to propose an approximate top-k selection algorithm, which is computing-efficient for GPUs, to improve the scaling efficiency of TopK-SGD by significantly reducing the computing overhead.
****
Entropy Penalty: Towards Generalization Beyond the IID Assumption
Devansh Arpit,Caiming Xiong,Richard Socher
It has been shown that instead of learning actual object features, deep networks tend to exploit non-robust (spurious) discriminative features that are shared between training and test sets. Therefore, while they achieve state of the art performance on such test sets, they achieve poor generalization on out of distribution (OOD) samples where the IID (independent, identical distribution) assumption breaks and the distribution of non-robust features shifts. Through theoretical and empirical analysis, we show that this happens because maximum likelihood training (without appropriate regularization) leads the model to depend on all the correlations (including spurious ones) present between inputs and targets in the dataset. We then show evidence that the information bottleneck (IB) principle can address this problem. To do so, we propose a regularization approach based on IB called Entropy Penalty, that reduces the model's dependence on spurious features-- features corresponding to such spurious correlations. This allows deep networks trained with Entropy Penalty to generalize well even under distribution shift of spurious features. As a controlled test-bed for evaluating our claim, we train deep networks with Entropy Penalty on a colored MNIST (C-MNIST) dataset and show that it is able to generalize well on vanilla MNIST, MNIST-M and SVHN datasets in addition to an OOD version of C-MNIST itself. The baseline regularization methods we compare against fail to generalize on this test-bed.

****
Improving Semantic Parsing with Neural Generator-Reranker Architecture
Huseyin A. Inan,Gaurav Singh Tomar,Huapu Pan
Semantic parsing is the problem of deriving machine interpretable meaning repres entations from natural language utterances. Neural models with encoder-decoder a rchitectures have recently achieved substantial improvements over traditional me thods. Although neural semantic parsers appear to have relatively high recall us ing large beam sizes, there is room for improvement with respect to one-best pre cision. In this work, we propose a generator-reranker architecture for semantic parsing. The generator produces a list of potential candidates and the reranker, which consists of a pre-processing step for the candidates followed by a novel critic network, reranks these candidates based on the similarity between each ca ndidate and the input sentence. We show the advantages of this approach along wi th how it improves the parsing performance through extensive analysis. We experi ment our model on three semantic parsing datasets (GEO, ATIS, and OVERNIGHT). Th e overall architecture achieves the state-of-the-art results in all three datase ts.
****
Learning a Behavioral Repertoire from Demonstrations
Niels Justesen,Miguel González Duque,Daniel Cabarcas Jaramillo,Jean-Baptiste Mou ret,Sebastian Risi
Imitation Learning (IL) is a machine learning approach to learn a policy from a set of demonstrations. IL can be useful to kick-start learning before applying r einforcement learning (RL) but it can also be useful on its own, e.g. to learn t o imitate human players in video games. However, a major limitation of current I L approaches is that they learn only a single ``"average" policy based on a data set that possibly contains demonstrations of numerous different types of behavio rs. In this paper, we present a new approach called Behavioral Repertoire Imitat ion Learning (BRIL) that instead learns a repertoire of behaviors from a set of demonstrations by augmenting the state-action pairs with behavioral descriptions . The outcome of this approach is a single neural network policy conditioned on a behavior description that can be precisely modulated. We apply this approach t o train a policy on 7,777 human demonstrations for the build-order planning task in StarCraft II. Dimensionality reduction techniques are applied to construct a low-dimensional behavioral space from the high-dimensional army unit compositio n of each demonstration. The results demonstrate that the learned policy can be effectively manipulated to express distinct behaviors. Additionally, by applying the UCB1 algorithm, the policy can adapt its behavior -in-between games- to rea ch a performance beyond that of the traditional IL baseline approach.
****
GRAPH NEIGHBORHOOD ATTENTIVE POOLING
Zekarias Tilahun Kefato,Sarunas Girdzijauskas
Network representation learning (NRL) is a powerful technique for learning low-d imensional vector representation of high-dimensional and sparse graphs. Most stu dies explore the structure and meta data associated with the graph using random walks and employ a unsupervised or semi-supervised learning schemes. Learning in these methods is context-free, because only a single representation per node is learned. Recently studies have argued on the sufficiency of a single representa tion and proposed a context-sensitive approach that proved to be highly effectiv e in applications such as link prediction and ranking.
However, most of these methods rely on additional textual features that require RNNs or CNNs to capture high-level features or rely on a community detection alg orithm to identifying multiple contexts of a node.
In this study, without requiring additional features nor a community detection a lgorithm, we propose a novel context-sensitive algorithm called GAP that learns to attend on different part of a node’s neighborhood using attentive pooling net works. We show the efficacy of GAP using three real-world datasets on link predi ction and node clustering tasks and compare it against 10 popular and state-of-t he-art (SOTA) baselines. GAP consistently outperforms them and achieves up to ≈9 % and ≈20% gain over the best performing methods on link prediction and clusteri

ng tasks, respectively.
****
Deep symbolic regression
Brenden K. Petersen
Discovering the underlying mathematical expressions describing a dataset is a co re challenge for artificial intelligence. This is the problem of symbolic regres sion. Despite recent advances in training neural networks to solve complex tasks , deep learning approaches to symbolic regression are lacking. We propose a fram ework that combines deep learning with symbolic regression via a simple idea: us e a large model to search the space of small models. More specifically, we use a recurrent neural network to emit a distribution over tractable mathematical exp ressions, and employ reinforcement learning to train the network to generate bet ter-fitting expressions. Our algorithm significantly outperforms standard geneti c programming-based symbolic regression in its ability to exactly recover symbol ic expressions on a series of benchmark problems, both with and without added no ise. More broadly, our contributions include a framework that can be applied to optimize hierarchical, variable-length objects under a black-box performance met ric, with the ability to incorporate a priori constraints in situ.
****
Autoencoders and Generative Adversarial Networks for Imbalanced Sequence Classif ication
Stephanie Ger,Diego Klabjan
We introduce a novel synthetic oversampling method for variable length, multi- f eature sequence datasets based on autoencoders and generative adversarial net- w orks. We show that this method improves classification accuracy for highly imbal anced sequence classification tasks. We show that this method outperforms standa rd oversampling techniques that use techniques such as SMOTE and autoencoders. W e also use generative adversarial networks on the majority class as an outlier d etection method for novelty detection, with limited classification improvement. We show that the use of generative adversarial network based synthetic data impr oves classification model performance on a variety of sequence data sets.

****
Uncertainty-Aware Prediction for Graph Neural Networks
Xujiang Zhao,Feng Chen,Shu Hu,jin-Hee Cho
Thanks to graph neural networks (GNNs), semi-supervised node classification has shown the state-of-the-art performance in graph data. However, GNNs do not cons ider any types of uncertainties associated with the class probabilities to minim ize risk due to misclassification under uncertainty in real life. In this work, we propose a Bayesian deep learning framework reflecting various types of uncert ainties for classification predictions by leveraging the powerful modeling and l earning capabilities of GNNs. We considered multiple uncertainty types in both d eep learning (DL) and belief/evidence theory domains. We treat the predictions o f a Bayesian GNN (BGNN) as nodes' multinomial subjective opinions in a graph bas ed on Dirichlet distributions where each belief mass is a belief probability of each class. By collecting evidence from the given labels of training nodes, the BGNN model is designed for accurately predicting probabilities of each class and detecting out-of-distribution. We validated the outperformance of the proposed BGNN, compared to the state-of-the-art counterparts in terms of the accuracy of node classification prediction and out-of-distribution detection based on six r eal network datasets.
****
Training Deep Neural Networks by optimizing over nonlocal paths in hyperparamete r space
Vlad Pushkarov,Yonathan Efroni,Mykola Maksymenko,Maciej Koch-Janusz
Hyperparameter optimization is both a practical issue and an interesting theoret ical problem in training of deep architectures. Despite many recent advances the most commonly used methods almost universally involve training multiple and dec oupled copies of the model, in effect sampling the hyperparameter space. We show that at a negligible additional computational cost, results can be improved by

sampling \emph{nonlocal paths} instead of points in hyperparameter space. To thi
s end we interpret hyperparameters as controlling the level of correlated noise
in training, which can be mapped to an effective temperature. The usually indepe
ndent instances of the model are coupled and allowed to exchange their hyperpara
meters throughout the training using the well established parallel tempering tec
hnique of statistical physics. Each simulation corresponds then to a unique path
, or history, in the joint hyperparameter/model-parameter space. We provide empi
rical tests of our method, in particular for dropout and learning rate optimizat
ion. We observed faster training and improved resistance to overfitting and show
ed a systematic decrease in the absolute validation error, improving over benchm
ark results.
****
Lattice Representation Learning
Luis A Lastras
We introduce the notion of \emph{lattice representation learning}, in which the
representation for some object of interest (e.g. a sentence or an image) is a la
ttice point in an Euclidean space. Our main contribution is a result for replaci
ng an objective function which employs lattice quantization with an objective fu
nction in which quantization is absent, thus allowing optimization techniques ba
sed on gradient descent to apply; we call the resulting algorithms \emph{dithere
d stochastic gradient descent} algorithms as they are designed explicitly to all
ow for an optimization procedure where only local information is employed. We al
so argue that a technique commonly used in Variational Auto-Encoders (Gaussian p
riors and Gaussian approximate posteriors) is tightly connected with the idea of
 lattice representations, as the quantization error in good high dimensional lat
tices can be modeled as a Gaussian distribution. We use a traditional encoder/de
coder architecture to explore the idea of latticed valued representations, and p
rovide experimental evidence of the potential of using lattice representations b
y modifying the \texttt{OpenNMT-py} generic \texttt{seq2seq} architecture so th
at it can implement not only Gaussian dithering of representations, but also the
 well known straight-through estimator and its application to vector quantizatio
n.

****
Omnibus Dropout for Improving The Probabilistic Classification Outputs of ConvNe
ts
Zhilu Zhang,Adrian V. Dalca,Mert R. Sabuncu
While neural network models achieve impressive classification accuracy across di
fferent tasks, they can suffer from poor calibration of their probabilistic pred
ictions. A Bayesian perspective has recently suggested that dropout, a regulariz
ation strategy popularly used during training, can be employed to obtain better
probabilistic predictions at test time (Gal & Ghahramani, 2016a). However, empir
ical results so far have not been encouraging, particularly with convolutional n
etworks. In this paper, through the lens of ensemble learning, we associate this
 unsatisfactory performance with the correlation between the models sampled with
 dropout. Motivated by this, we explore the use of various structured dropout te
chniques to promote model diversity and improve the quality of probabilistic pre
dictions. We also propose an omnibus dropout strategy that combines various stru
ctured dropout methods. Using the SVHN, CIFAR-10 and CIFAR-100 datasets, we empi
rically demonstrate the superior performance of omnibus dropout relative to seve
ral widely used strong baselines in addition to regular dropout. Lastly, we show
 the merit of omnibus dropout in a Bayesian active learning application.
****
Deep Multiple Instance Learning for Taxonomic Classification of Metagenomic read
 sets
Andreas Georgiou,Vincent Fortuin,Harun Mustafa,Gunnar Rätsch
Metagenomic studies have increasingly utilized sequencing technologies in order
to analyze DNA fragments found in environmental samples. It can provide useful i
nsights for studying the interactions between hosts and microbes, infectious dis
ease proliferation, and novel species discovery. One important step in this anal

ysis is the taxonomic classification of those DNA fragments. Of particular inter
est is the determination of the distribution of the taxa of microbes in metageno
mic samples. Recent attempts using deep learning focus on architectures that cla
ssify single DNA reads independently from each other. In this work, we attempt t
o solve the task of directly predicting the distribution over the taxa of whole
metagenomic read sets. We formulate this task as a Multiple Instance Learning (M
IL) problem. We extend architectures used in single-read taxonomic classificatio
n with two different types of permutation-invariant MIL pooling layers: a) deeps
ets and b) attention-based pooling. We illustrate that our architecture can expl
oit the co-occurrence of species in metagenomic read sets and outperforms the si
ngle-read architectures in predicting the distribution over the taxa at higher t
axonomic ranks.
****

Budgeted Training: Rethinking Deep Neural Network Training Under Resource Constr
aints

Mengtian Li,Ersin Yumer,Deva Ramanan

In most practical settings and theoretical analyses, one assumes that a model ca
n be trained until convergence. However, the growing complexity of machine learn
ing datasets and models may violate such assumptions. Indeed, current approaches
 for hyper-parameter tuning and neural architecture search tend to be limited by
 practical resource constraints. Therefore, we introduce a formal setting for st
udying training under the non-asymptotic, resource-constrained regime, i.e., bud
geted training. We analyze the following problem: "given a dataset, algorithm, a
nd fixed resource budget, what is the best achievable performance?" We focus on
the number of optimization iterations as the representative resource. Under such
 a setting, we show that it is critical to adjust the learning rate schedule acc
ording to the given budget. Among budget-aware learning schedules, we find simpl
e linear decay to be both robust and high-performing. We support our claim throu
gh extensive experiments with state-of-the-art models on ImageNet (image classif
ication), Kinetics (video classification), MS COCO (object detection and instanc
e segmentation), and Cityscapes (semantic segmentation). We also analyze our res
ults and find that the key to a good schedule is budgeted convergence, a phenome
non whereby the gradient vanishes at the end of each allowed budget. We also rev
isit existing approaches for fast convergence and show that budget-aware learnin
g schedules readily outperform such approaches under (the practical but under-ex
plored) budgeted training setting.
****

RoBERTa: A Robustly Optimized BERT Pretraining Approach

Yinhan Liu,Myle Ott,Naman Goyal,Jingfei Du,Mandar Joshi,Danqi Chen,Omer Levy,Mik
e Lewis,Luke Zettlemoyer,Veselin Stoyanov

Language model pretraining has led to significant performance gains but careful
comparison between different approaches is challenging. Training is computationa
lly expensive, often done on private datasets of different sizes, and, as we wil
l show, hyperparameter choices have significant impact on the final results. We
present a replication study of BERT pretraining (Devlin et al., 2019) that caref
ully measures the impact of many key hyperparameters and training data size. We
find that BERT was significantly undertrained, and can match or exceed the perfo
rmance of every model published after it. Our best model achieves state-of-the-a
rt results on GLUE, RACE, SQuAD, SuperGLUE and XNLI. These results highlight the
 importance of previously overlooked design choices, and raise questions about t
he source of recently reported improvements. We release our models and code.
****

Deep Semi-Supervised Anomaly Detection

Lukas Ruff,Robert A. Vandermeulen,Nico Görnitz,Alexander Binder,Emmanuel Müller,
Klaus-Robert Müller,Marius Kloft

Deep approaches to anomaly detection have recently shown promising results over
shallow methods on large and complex datasets. Typically anomaly detection is tr
eated as an unsupervised learning problem. In practice however, one may have---i
n addition to a large set of unlabeled samples---access to a small pool of label
ed samples, e.g. a subset verified by some domain expert as being normal or anom

alous. Semi-supervised approaches to anomaly detection aim to utilize such label ed samples, but most proposed methods are limited to merely including labeled no rmal samples. Only a few methods take advantage of labeled anomalies, with exist ing deep approaches being domain-specific. In this work we present Deep SAD, an end-to-end deep methodology for general semi-supervised anomaly detection. We fu rther introduce an information-theoretic framework for deep anomaly detection ba sed on the idea that the entropy of the latent distribution for normal data shou ld be lower than the entropy of the anomalous distribution, which can serve as a theoretical interpretation for our method. In extensive experiments on MNIST, F ashion-MNIST, and CIFAR-10, along with other anomaly detection benchmark dataset s, we demonstrate that our method is on par or outperforms shallow, hybrid, and deep competitors, yielding appreciable performance improvements even when provid ed with only little labeled data.
****
GNN-FiLM: Graph Neural Networks with Feature-wise Linear Modulation
Marc Brockschmidt
This paper presents a new Graph Neural Network (GNN) type using feature-wise lin ear modulation (FiLM). Many standard GNN variants propagate information along th e edges of a graph by computing ``messages'' based only on the representation of the source of each edge. In GNN-FiLM, the representation of the target node of an edge is additionally used to compute a transformation that can be applied to all incoming messages, allowing feature-wise modulation of the passed informatio n.

Results of experiments comparing different GNN architectures on three tasks from the literature are presented, based on re-implementations of baseline methods. Hyperparameters for all methods were found using extensive search, yielding some what surprising results: differences between baseline models are smaller than re ported in the literature. Nonetheless, GNN-FiLM outperforms baseline methods on a regression task on molecular graphs and performs competitively on other tasks.

****
Out-of-distribution Detection in Few-shot Classification
Kuan-Chieh Wang,Paul Vicol,Eleni Triantafillou,Chia-Cheng Liu,Richard Zemel
In many real-world settings, a learning model must perform few-shot classificati on: learn to classify examples from unseen classes using only a few labeled exam ples per class.
Additionally, to be safely deployed, it should have the ability to detect out-of -distribution inputs: examples that do not belong to any of the classes.
While both few-shot classification and out-of-distribution detection are popular topics,
their combination has not been studied. In this work, we propose tasks for out-o f-distribution detection in the few-shot setting and establish benchmark dataset s, based on four popular few-shot classification datasets. Then, we propose two new methods for this task and investigate their performance.
In sum, we establish baseline out-of-distribution detection results using standa rd metrics on new benchmark datasets and show improved results with our proposed methods.
****
Mirror-Generative Neural Machine Translation
Zaixiang Zheng,Hao Zhou,Shujian Huang,Lei Li,Xin-Yu Dai,Jiajun Chen
Training neural machine translation models (NMT) requires a large amount of para llel corpus, which is scarce for many language pairs. However, raw non-parallel corpora are often easy to obtain. Existing approaches have not exploited the ful l potential of non-parallel bilingual data either in training or decoding. In th is paper, we propose the mirror-generative NMT (MGNMT), a single unified archite cture that simultaneously integrates the source to target translation model, the target to source translation model, and two language models. Both translation m odels and language models share the same latent semantic space, therefore both t ranslation directions can learn from non-parallel data more effectively. Besides

, the translation models and language models can collaborate together during dec
oding. Our experiments show that the proposed MGNMT consistently outperforms exi
sting approaches in a variety of scenarios and language pairs, including resourc
e-rich and low-resource situations.
****
Frustratingly easy quasi-multitask learning
Gábor Berend,Norbert Kis-Szabó
We propose the technique of quasi-multitask learning (Q-MTL), a simple and easy
to implement modification of standard multitask learning, in which the tasks to
 be modeled are identical. We illustrate it through a series of sequence labelin
g experiments over a diverse set of languages, that applying Q-MTL consistently
increases the generalization ability of the applied models. The proposed archite
cture can be regarded as a new regularization technique encouraging the model to
 develop an internal representation of the problem at hand that is beneficial to
 multiple output units of the classifier at the same time. This property hampers
 the convergence to such internal representations which are highly specific and
tailored for a classifier with a particular set of parameters. Our experiments c
orroborate that by relying on the proposed algorithm, we can approximate the qua
lity of an ensemble of classifiers at a fraction of computational resources requ
ired. Additionally, our results suggest that Q-MTL handles the presence of noisy
 training labels better than ensembles.

****
Interpreting video features: a comparison of 3D convolutional networks and convo
lutional LSTM networks
Joonatan Mänttäri*,Sofia Broomé*,John Folkesson,Hedvig Kjellström
A number of techniques for interpretability have been presented for deep learnin
g
in computer vision, typically with the goal of understanding what it is that the
 networks
have actually learned underneath a given classification decision. However,
when it comes to deep video architectures, interpretability is still in its infa
ncy and
we do not yet have a clear concept of how we should decode spatiotemporal featur
es.
In this paper, we present a study comparing how 3D convolutional networks
and convolutional LSTM networks respectively learn features across temporally
dependent frames. This is the first comparison of two video models that both
convolve to learn spatial features but that have principally different methods o
f
modeling time. Additionally, we extend the concept of meaningful perturbation
introduced by Fong & Vedaldi (2017) to the temporal dimension to search for the
most meaningful part of a sequence for a classification decision.
****
TrojanNet: Exposing the Danger of Trojan Horse Attack on Neural Networks
Chuan Guo,Ruihan Wu,Kilian Q. Weinberger
The complexity of large-scale neural networks can lead to poor understanding of
their internal details. We show that this opaqueness provides an opportunity fo
r adversaries to embed unintended functionalities into the network in the form o
f Trojan horse attacks. Our novel framework hides the existence of a malicious n
etwork within a benign transport network. Our attack is flexible, easy to execut
e, and difficult to detect. We prove theoretically that the malicious network's
detection is computationally infeasible and demonstrate empirically that the tra
nsport network does not compromise its disguise. Our attack exposes an important
, previously unknown loophole that unveils a new direction in machine learning s
ecurity.
****
Robust Learning with Jacobian Regularization
Judy Hoffman,Daniel A. Roberts,Sho Yaida
Design of reliable systems must guarantee stability against input perturbations.

In machine learning, such guarantee entails preventing overfitting and ensuring robustness of models against corruption of input data. In order to maximize stability, we analyze and develop a computationally efficient implementation of Jacobian regularization that increases classification margins of neural networks. The stabilizing effect of the Jacobian regularizer leads to significant improvements in robustness, as measured against both random and adversarial input perturbations, without severely degrading generalization properties on clean data.
****
Generalized Inner Loop Meta-Learning
Edward Grefenstette,Brandon Amos,Denis Yarats,Phu Mon Htut,Artem Molchanov,Franziska Meier,Douwe Kiela,Kyunghyun Cho,Soumith Chintala
Many (but not all) approaches self-qualifying as "meta-learning" in deep learning and reinforcement learning fit a common pattern of approximating the solution to a nested optimization problem. In this paper, we give a formalization of this shared pattern, which we call GIMLI, prove its general requirements, and derive a general-purpose algorithm for implementing similar approaches. Based on this analysis and algorithm, we describe a library of our design, unnamedlib, which we share with the community to assist and enable future research into these kinds of meta-learning approaches. We end the paper by showcasing the practical applications of this framework and library through illustrative experiments and ablation studies which they facilitate.
****
Sign Bits Are All You Need for Black-Box Attacks
Abdullah Al-Dujaili,Una-May O'Reilly
We present a novel black-box adversarial attack algorithm with state-of-the-art model evasion rates for query efficiency under ℓ_∞ and ℓ_2 metrics. It exploits a \textit{sign-based}, rather than magnitude-based, gradient estimation approach that shifts the gradient estimation from continuous to binary black-box optimization. It adaptively constructs queries to estimate the gradient, one query relying upon the previous, rather than re-estimating the gradient each step with random query construction. Its reliance on sign bits yields a smaller memory footprint and it requires neither hyperparameter tuning or dimensionality reduction. Further, its theoretical performance is guaranteed and it can characterize adversarial subspaces better than white-box gradient-aligned subspaces. On two public black-box attack challenges and a model robustly trained against transfer attacks, the algorithm's evasion rates surpass all submitted attacks. For a suite of published models, the algorithm is $3.8\times$ less failure-prone while spending $2.5\times$ fewer queries versus the best combination of state of art algorithms. For example, it evades a standard MNIST model using just 12 queries on average. Similar performance is observed on a standard IMAGENET model with an average of 579 queries.
****
Learning Hierarchical Discrete Linguistic Units from Visually-Grounded Speech
David Harwath*,Wei-Ning Hsu*,James Glass
In this paper, we present a method for learning discrete linguistic units by incorporating vector quantization layers into neural models of visually grounded speech. We show that our method is capable of capturing both word-level and sub-word units, depending on how it is configured. What differentiates this paper from prior work on speech unit learning is the choice of training objective. Rather than using a reconstruction-based loss, we use a discriminative, multimodal grounding objective which forces the learned units to be useful for semantic image retrieval. We evaluate the sub-word units on the ZeroSpeech 2019 challenge, achieving a 27.3% reduction in ABX error rate over the top-performing submission, while keeping the bitrate approximately the same. We also present experiments demonstrating the noise robustness of these units. Finally, we show that a model with multiple quantizers can simultaneously learn phone-like detectors at a lower layer and word-like detectors at a higher layer. We show that these detectors are highly accurate, discovering 279 words with an F1 score of greater than 0.5.
****
Pre-training as Batch Meta Reinforcement Learning with tiMe

Quan Vuong,Shuang Liu,Minghua Liu,Kamil Ciosek,Hao Su,Henrik Iskov Christensen
Pre-training is transformative in supervised learning: a large network trained with large and existing datasets can be used as an initialization when learning a new task. Such initialization speeds up convergence and leads to higher performance. In this paper, we seek to understand what the formalization for pre-training from only existing and observational data in Reinforcement Learning (RL) is and whether it is possible. We formulate the setting as Batch Meta Reinforcement Learning. We identify MDP mis-identification to be a central challenge and motivate it with theoretical analysis. Combining ideas from Batch RL and Meta RL, we propose tiMe, which learns distillation of multiple value functions and MDP embeddings from only existing data. In challenging control tasks and without fine-tuning on unseen MDPs, tiMe is competitive with state-of-the-art model-free RL method trained with hundreds of thousands of environment interactions.
****
Reinforced active learning for image segmentation
Arantxa Casanova,Pedro O. Pinheiro,Negar Rostamzadeh,Christopher J. Pal
Learning-based approaches for semantic segmentation have two inherent challenges. First, acquiring pixel-wise labels is expensive and time-consuming. Second, realistic segmentation datasets are highly unbalanced: some categories are much more abundant than others, biasing the performance to the most represented ones. In this paper, we are interested in focusing human labelling effort on a small subset of a larger pool of data, minimizing this effort while maximizing performance of a segmentation model on a hold-out set. We present a new active learning strategy for semantic segmentation based on deep reinforcement learning (RL). An agent learns a policy to select a subset of small informative image regions -- opposed to entire images -- to be labeled, from a pool of unlabeled data. The region selection decision is made based on predictions and uncertainties of the segmentation model being trained. Our method proposes a new modification of the deep Q-network (DQN) formulation for active learning, adapting it to the large-scale nature of semantic segmentation problems. We test the proof of concept in CamVid and provide results in the large-scale dataset Cityscapes. On Cityscapes, our deep RL region-based DQN approach requires roughly 30% less additional labeled data than our most competitive baseline to reach the same performance. Moreover, we find that our method asks for more labels of under-represented categories compared to the baselines, improving their performance and helping to mitigate class imbalance.
****
Neural Architecture Search by Learning Action Space for Monte Carlo Tree Search
Linnan Wang,Saining Xie,Teng Li,Rodrigo Fonseca,Yuandong Tian
Neural Architecture Search (NAS) has emerged as a promising technique for automatic neural network design. However, existing NAS approaches often utilize manually designed action space, which is not directly related to the performance metric to be optimized (e.g., accuracy). As a result, using manually designed action space to perform NAS often leads to sample-inefficient explorations of architectures and thus can be sub-optimal. In order to improve sample efficiency, this paper proposes Latent Action Neural Architecture Search (LaNAS) that learns actions to recursively partition the search space into good or bad regions that contain networks with concentrated performance metrics, i.e., low variance. During the search phase, as different architecture search action sequences lead to regions of different performance, the search efficiency can be significantly improved by biasing towards the good regions. On the largest NAS dataset NASBench-101, our experimental results demonstrated that LaNAS is 22x, 14.6x, 12.4x, 6.8x, 16.5x more sample-efficient than Random Search, Regularized Evolution, Monte Carlo Tree Search, Neural Architecture Optimization, and Bayesian Optimization, respectively. When applied to the open domain, LaNAS achieves 98.0% accuracy on CIFAR-10 and 75.0% top1 accuracy on ImageNet in only 803 samples, outperforming SOTA AmoebaNet with 33x fewer samples.
****
Gradientless Descent: High-Dimensional Zeroth-Order Optimization
Daniel Golovin,John Karro,Greg Kochanski,Chansoo Lee,Xingyou Song,Qiuyi Zhang

Zeroth-order optimization is the process of minimizing an objective $f(x)$, give
n oracle access to evaluations at adaptively chosen inputs x. In this paper, w
e present two simple yet powerful GradientLess Descent (GLD) algorithms that do
not rely on an underlying gradient estimate and are numerically stable. We analy
ze our algorithm from a novel geometric perspective and we show that for {\it an
y monotone transform} of a smooth and strongly convex objective with latent dime
nsion $k \ge n$, we present a novel analysis that shows convergence within an $\
epsilon$-ball of the optimum in $O(kQ\log(n)\log(R/\epsilon))$ evaluations, wher
e the input dimension is n, R is the diameter of the input space and Q is
the condition number. Our rates are the first of its kind to be both 1) poly-log
arithmically dependent on dimensionality and 2) invariant under monotone transfo
rmations. We further leverage our geometric perspective to show that our analysi
s is optimal. Both monotone invariance and its ability to utilize a low latent d
imensionality are key to the empirical success of our algorithms, as demonstrate
d on synthetic and MuJoCo benchmarks.

****
Equivariant Entity-Relationship Networks
Devon Graham,Siamak Ravanbakhsh
Due to its extensive use in databases, the relational model is ubiquitous in rep
resenting big-data. However, recent progress in deep learning with relational da
ta has been focused on (knowledge) graphs. In this paper we propose Equivariant
Entity-Relationship Networks, the class of parameter-sharing neural networks der
ived from the entity-relationship model. We prove that our proposed feed-forward
 layer is the most expressive linear layer under the given equivariance constrai
nts, and subsumes recently introduced equivariant models for sets, exchangeable
tensors, and graphs. The proposed feed-forward layer has linear complexity in th
e the data and can be used for both inductive and transductive reasoning about r
elational databases, including database embedding, and the prediction of missing
 records. This, provides a principled theoretical foundation for the application
 of deep learning to one of the most abundant forms of data.
****
Modeling Fake News in Social Networks with Deep Multi-Agent Reinforcement Learni
ng
Christoph Aymanns,Matthias Weber,Co-Pierre Georg,Jakob Foerster
We develop a practical and flexible computational model of fake news on social n
etworks in which agents act according to learned best response functions. We ach
ieve this by extending an information aggregation game to allow for fake news an
d by representing agents as recurrent deep Q-networks (DQN) trained by independe
nt Q-learning. In the game, agents repeatedly guess whether a claim is true or f
alse taking into account an informative private signal and observations of actio
ns of their neighbors on the social network in the previous period. We incorpora
te fake news into the model by adding an adversarial agent, the attacker, that e
ither provides biased private signals to or takes over a subset of agents. The a
ttacker can follow either a hand-tuned or trained policy. Our model allows us to
 tackle questions that are analytically intractable in fully rational models, wh
ile ensuring that agents follow reasonable best response functions. Our results
highlight the importance of awareness, privacy and social connectivity in curbin
g the adverse effects of fake news.

****
On the "steerability" of generative adversarial networks
Ali Jahanian*,Lucy Chai*,Phillip Isola
An open secret in contemporary machine learning is that many models work beautif
ully on standard benchmarks but fail to generalize outside the lab. This has bee
n attributed to biased training data, which provide poor coverage over real worl
d events. Generative models are no exception, but recent advances in generative
adversarial networks (GANs) suggest otherwise -- these models can now synthesize
 strikingly realistic and diverse images. Is generative modeling of photos a sol
ved problem? We show that although current GANs can fit standard datasets very w

ell, they still fall short of being comprehensive models of the visual manifold. In particular, we study their ability to fit simple transformations such as camera movements and color changes. We find that the models reflect the biases of the datasets on which they are trained (e.g., centered objects), but that they also exhibit some capacity for generalization: by "steering" in latent space, we can shift the distribution while still creating realistic images. We hypothesize that the degree of distributional shift is related to the breadth of the training data distribution. Thus, we conduct experiments to quantify the limits of GAN transformations and introduce techniques to mitigate the problem. Code is released on our project page: https://ali-design.github.io/gan_steerability/
****

Improving Differentially Private Models with Active Learning
Zhengli Zhao,Nicolas Papernot,Sameer Singh,Neoklis Polyzotis,Augustus Odena
Broad adoption of machine learning techniques has increased privacy concerns for models trained on sensitive data such as medical records. Existing techniques for training differentially private (DP) models give rigorous privacy guarantees, but applying these techniques to neural networks can severely degrade model performance. This performance reduction is an obstacle to deploying private models in the real world. In this work, we improve the performance of DP models by fine-tuning them through active learning on public data. We introduce two new techniques - DiversePublic and NearPrivate - for doing this fine-tuning in a privacy-aware way. For the MNIST and SVHN datasets, these techniques improve state-of-the-art accuracy for DP models while retaining privacy guarantees.
****

Matrix Multilayer Perceptron
Jalil Taghia,Maria Bånkestad,Fredrik Lindsten,Thomas Schön
Models that output a vector of responses given some inputs, in the form of a conditional mean vector, are at the core of machine learning. This includes neural networks such as the multilayer perceptron (MLP). However, models that output a symmetric positive definite (SPD) matrix of responses given inputs, in the form of a conditional covariance function, are far less studied, especially within the context of neural networks. Here, we introduce a new variant of the MLP, referred to as the matrix MLP, that is specialized at learning SPD matrices. Our construction not only respects the SPD constraint, but also makes explicit use of it. This translates into a model which effectively performs the task of SPD matrix learning even in scenarios where data are scarce. We present an application of the model in heteroscedastic multivariate regression, including convincing performance on six real-world datasets.
****

Feature-Robustness, Flatness and Generalization Error for Deep Neural Networks
Henning Petzka,Linara Adilova,Michael Kamp,Cristian Sminchisescu
The performance of deep neural networks is often attributed to their automated, task-related feature construction. It remains an open question, though, why this leads to solutions with good generalization, even in cases where the number of parameters is larger than the number of samples. Back in the 90s, Hochreiter and Schmidhuber observed that flatness of the loss surface around a local minimum correlates with low generalization error. For several flatness measures, this correlation has been empirically validated. However, it has recently been shown that existing measures of flatness cannot theoretically be related to generalization: if a network uses ReLU activations, the network function can be reparameterized without changing its output in such a way that flatness is changed almost arbitrarily. This paper proposes a natural modification of existing flatness measures that results in invariance to reparameterization. The proposed measures imply a robustness of the network to changes in the input and the hidden layers. Connecting this feature robustness to generalization leads to a generalized definition of the representativeness of data. With this, the generalization error of a model trained on representative data can be bounded by its feature robustness which depends on our novel flatness measure.
****

TriMap: Large-scale Dimensionality Reduction Using Triplets

Ehsan Amid,Manfred K. Warmuth
We introduce ``TriMap''; a dimensionality reduction technique based on triplet c onstraints that preserves the global accuracy of the data better than the other commonly used methods such as t-SNE, LargeVis, and UMAP. To quantify the global accuracy, we introduce a score which roughly reflects the relative placement of the clusters rather than the individual points. We empirically show the excelle nt performance of TriMap on a large variety of datasets in terms of the quality of the embedding as well as the runtime. On our performance benchmarks, TriMap e asily scales to millions of points without depleting the memory and clearly outp erforms t-SNE, LargeVis, and UMAP in terms of runtime.
****

LEARNED STEP SIZE QUANTIZATION
Steven K. Esser,Jeffrey L. McKinstry,Deepika Bablani,Rathinakumar Appuswamy,Dhar mendra S. Modha
Deep networks run with low precision operations at inference time offer power an d space advantages over high precision alternatives, but need to overcome the ch allenge of maintaining high accuracy as precision decreases. Here, we present a method for training such networks, Learned Step Size Quantization, that achieves the highest accuracy to date on the ImageNet dataset when using models, from a variety of architectures, with weights and activations quantized to 2-, 3- or 4- bits of precision, and that can train 3-bit models that reach full precision bas eline accuracy. Our approach builds upon existing methods for learning weights i n quantized networks by improving how the quantizer itself is configured. Specif ically, we introduce a novel means to estimate and scale the task loss gradient at each weight and activation layer's quantizer step size, such that it can be l earned in conjunction with other network parameters. This approach works using d ifferent levels of precision as needed for a given system and requires only a si mple modification of existing training code.
****

Learning General and Reusable Features via Racecar-Training
You Xie,Nils Thuerey
We propose a novel training approach for improving the learning of generalizing features in neural networks. We augment the network with a reverse pass which ai ms for reconstructing the full sequence of internal states of the network. Despi te being a surprisingly simple change, we demonstrate that this forward-backward training approach, i.e. racecar training, leads to significantly more general f eatures to be extracted from a given data set. We demonstrate in our paper that a network obtained in this way is continually trained for the original task, it outperforms baseline models trained in a regular fashion. This improved performa nce is visible for a wide range of learning tasks from classification, to regres sion and stylization. In addition, networks trained with our approach exhibit im proved performance for task transfers. We additionally analyze the mutual inform ation of our networks to explain the improved generalizing capabilities.
****

Estimating counterfactual treatment outcomes over time through adversarially bal anced representations
Ioana Bica,Ahmed M Alaa,James Jordon,Mihaela van der Schaar
Identifying when to give treatments to patients and how to select among multiple treatments over time are important medical problems with a few existing solutio ns. In this paper, we introduce the Counterfactual Recurrent Network (CRN), a no vel sequence-to-sequence model that leverages the increasingly available patient observational data to estimate treatment effects over time and answer such medi cal questions. To handle the bias from time-varying confounders, covariates affe cting the treatment assignment policy in the observational data, CRN uses domain adversarial training to build balancing representations of the patient history. At each timestep, CRN constructs a treatment invariant representation which rem oves the association between patient history and treatment assignments and thus can be reliably used for making counterfactual predictions. On a simulated model of tumour growth, with varying degree of time-dependent confounding, we show ho w our model achieves lower error in estimating counterfactuals and in choosing t

he correct treatment and timing of treatment than current state-of-the-art metho
ds.
****

Poincaré Wasserstein Autoencoder
Ivan Ovinnikov
This work presents the Poincaré Wasserstein Autoencoder, a reformulation of
the recently proposed Wasserstein autoencoder framework on a non-Euclidean
manifold, the Poincaré ball model of the hyperbolic space H n . By assuming the
latent space to be hyperbolic, we can use its intrinsic hierarchy to impose stru
cture
on the learned latent space representations. We show that for datasets with late
nt
hierarchies, we can recover the structure in a low-dimensional latent space. We
also demonstrate the model in the visual domain to analyze some of its propertie
s
and show competitive results on a graph link prediction task.
****

Robust Instruction-Following in a Situated Agent via Transfer-Learning from Text
Felix Hill,Sona Mokra,Nathaniel Wong,Tim Harley
Recent work has described neural-network-based agents that are trained to execut
e language-like commands in simulated worlds, as a step towards an intelligent a
gent or robot that can be instructed by human users. However, the instructions t
hat such agents are trained to follow are typically generated from templates (by
 an environment simulator), and do not reflect the varied or ambiguous expressio
ns used by real people. We address this issue by integrating language encoders t
hat are pretrained on large text corpora into a situated, instruction-following
agent. In a procedurally-randomized first-person 3D world, we first train agents
 to follow synthetic instructions requiring the identification, manipulation and
 relative positioning of visually-realistic objects models. We then show how the
se abilities can transfer to a context where humans provide instructions in natu
ral language, but only when agents are endowed with language encoding components
 that were pretrained on text-data. We explore techniques for integrating text-t
rained and environment-trained components into an agent, observing clear advanta
ges for the fully-contextual phrase representations computed by the well-known B
ERT model, and additional gains by integrating a self-attention operation optimi
zed to adapt BERT's representations for the agent's tasks and environment. These
  results bridge the gap between two successful strands of recent AI research: a
gent-centric behavior optimization and text-based representation learning.
****

Stochastic Conditional Generative Networks with Basis Decomposition
Ze Wang,Xiuyuan Cheng,Guillermo Sapiro,Qiang Qiu
While generative adversarial networks (GANs) have revolutionized machine learnin
g, a number of open questions remain to fully understand them and exploit their
power. One of these questions is how to efficiently achieve proper diversity and
 sampling of the multi-mode data space. To address this, we introduce BasisGAN,
a stochastic conditional multi-mode image generator. By exploiting the observati
on that a convolutional filter can be well approximated as a linear combination
of a small set of basis elements, we learn a plug-and-played basis generator to
stochastically generate basis elements, with just a few hundred of parameters, t
o fully embed stochasticity into convolutional filters. By sampling basis elemen
ts instead of filters, we dramatically reduce the cost of modeling the parameter
 space with no sacrifice on either image diversity or fidelity. To illustrate th
is proposed plug-and-play framework, we construct variants of BasisGAN based on
state-of-the-art conditional image generation networks, and train the networks b
y simply plugging in a basis generator, without additional auxiliary components,
 hyperparameters, or training objectives. The experimental success is complement
ed with theoretical results indicating how the perturbations introduced by the p
roposed sampling of basis elements can propagate to the appearance of generated
images.
****

Task-Based Top-Down Modulation Network for Multi-Task-Learning Applications
Hila Levi,Shimon Ullman
A general problem that received considerable recent attention is how to perform multiple tasks in the same network, maximizing both efficiency and prediction accuracy. A popular approach consists of a multi-branch architecture on top of a shared backbone, jointly trained on a weighted sum of losses. However, in many cases, the shared representation results in non-optimal performance, mainly due to an interference between conflicting gradients of uncorrelated tasks. Recent approaches address this problem by a channel-wise modulation of the feature-maps along the shared backbone, with task specific vectors, manually or dynamically tuned. Taking this approach a step further, we propose a novel architecture which modulate the recognition network channel-wise, as well as spatial-wise, with an efficient top-down image-dependent computation scheme. Our architecture uses no task-specific branches, nor task specific modules. Instead, it uses a top-down modulation network that is shared between all of the tasks. We show the effectiveness of our scheme by achieving on par or better results than alternative approaches on both correlated and uncorrelated sets of tasks. We also demonstrate our advantages in terms of model size, the addition of novel tasks and interpretability.
Code will be released.
****
Tensor Graph Convolutional Networks for Prediction on Dynamic Graphs
Osman Asif Malik,Shashanka Ubaru,Lior Horesh,Misha E. Kilmer,Haim Avron
Many irregular domains such as social networks, financial transactions, neuron connections, and natural language structures are represented as graphs. In recent years, a variety of graph neural networks (GNNs) have been successfully applied for representation learning and prediction on such graphs. However, in many of the applications, the underlying graph changes over time and existing GNNs are inadequate for handling such dynamic graphs. In this paper we propose a novel technique for learning embeddings of dynamic graphs based on a tensor algebra framework. Our method extends the popular graph convolutional network (GCN) for learning representations of dynamic graphs using the recently proposed tensor M-product technique. Theoretical results that establish the connection between the proposed tensor approach and spectral convolution of tensors are developed. Numerical experiments on real datasets demonstrate the usefulness of the proposed method for an edge classification task on dynamic graphs.
****
GraphNVP: an Invertible Flow-based Model for Generating Molecular Graphs
Kaushalya Madhawa,Katsuhiko Ishiguro,Kosuke Nakago,Motoki Abe
We propose GraphNVP, an invertible flow-based molecular graph generation model. Existing flow-based models only handle node attributes of a graph with invertible maps. In contrast, our model is the first invertible model for the whole graph components: both of dequantized node attributes and adjacency tensor are converted into latent vectors through two novel invertible flows. This decomposition yields the exact likelihood maximization on graph-structured data. We decompose the generation of a graph into two steps: generation of (i) an adjacency tensor and(ii) node attributes. We empirically demonstrate that our model and the two-step generation efficiently generates valid molecular graphs with almost no duplicated molecules, although there are no domain-specific heuristics ingrained in the model. We also confirm that the sampling (generation) of graphs is faster in magnitude than other models in our implementation. In addition, we observe that the learned latent space can be used to generate molecules with desired chemical properties
****
Language GANs Falling Short
Massimo Caccia,Lucas Caccia,William Fedus,Hugo Larochelle,Joelle Pineau,Laurent Charlin
Traditional natural language generation (NLG) models are trained using maximum likelihood estimation (MLE) which differs from the sample generation inference procedure. During training the ground truth tokens are passed to the model, howev

er, during inference, the model instead reads its previously generated samples - a phenomenon coined exposure bias. Exposure bias was hypothesized to be a root cause of poor sample quality and thus many generative adversarial networks (GANs) were proposed as a remedy since they have identical training and inference. H owever, many of the ensuing GAN variants validated sample quality improvements b ut ignored loss of sample diversity. This work reiterates the fallacy of quality -only metrics and clearly demonstrate that the well-established technique of red ucing softmax temperature can outperform GANs on a quality-only metric. Further, we establish a definitive quality-diversity evaluation procedure using temperat ure tuning over local and global sample metrics. Under this, we find that MLE mo dels consistently outperform the proposed GAN variants over the whole quality-di versity space. Specifically, we find that 1) exposure bias appears to be less o f an issue than the complications arising from non-differentiable, sequential GA N training; 2) MLE trained models provide a better quality/diversity trade-off compared to their GAN counterparts, all while being easier to train, easier to c ross-validate, and less computationally expensive.
****

GENESIS: Generative Scene Inference and Sampling with Object-Centric Latent Repr esentations
Martin Engelcke,Adam R. Kosiorek,Oiwi Parker Jones,Ingmar Posner
Generative latent-variable models are emerging as promising tools in robotics an d reinforcement learning. Yet, even though tasks in these domains typically invo lve distinct objects, most state-of-the-art generative models do not explicitly capture the compositional nature of visual scenes. Two recent exceptions, MONet and IODINE, decompose scenes into objects in an unsupervised fashion. Their unde rlying generative processes, however, do not account for component interactions. Hence, neither of them allows for principled sampling of novel scenes. Here we present GENESIS, the first object-centric generative model of 3D visual scenes capable of both decomposing and generating scenes by capturing relationships bet ween scene components. GENESIS parameterises a spatial GMM over images which is decoded from a set of object-centric latent variables that are either inferred s equentially in an amortised fashion or sampled from an autoregressive prior. We train GENESIS on several publicly available datasets and evaluate its performanc e on scene generation, decomposition, and semi-supervised learning.
****

Last-iterate convergence rates for min-max optimization
Jacob Abernethy,Kevin A. Lai,Andre Wibisono
While classic work in convex-concave min-max optimization relies on average-iter ate convergence results, the emergence of nonconvex applications such as trainin g Generative Adversarial Networks has led to renewed interest in last-iterate co nvergence guarantees. Proving last-iterate convergence is challenging because ma ny natural algorithms, such as Simultaneous Gradient Descent/Ascent, provably di verge or cycle even in simple convex-concave min-max settings, and previous work on global last-iterate convergence rates has been limited to the bilinear and c onvex-strongly concave settings. In this work, we show that the Hamiltonian Grad ient Descent (HGD) algorithm achieves linear convergence in a variety of more ge neral settings, including convex-concave problems that satisfy a “sufficiently b ilinear” condition. We also prove similar convergence rates for some parameter s ettings of the Consensus Optimization (CO) algorithm of Mescheder et al. 2017.
****

Poisoning Attacks with Generative Adversarial Nets
Luis Muñoz-González,Bjarne Pfitzner,Matteo Russo,Javier Carnerero-Cano,Emil C. L upu
Machine learning algorithms are vulnerable to poisoning attacks: An adversary ca n inject malicious points in the training dataset to influence the learning proc ess and degrade the algorithm's performance. Optimal poisoning attacks have alre ady been proposed to evaluate worst-case scenarios, modelling attacks as a bi-le vel optimization problem. Solving these problems is computationally demanding an d has limited applicability for some models such as deep networks. In this paper we introduce a novel generative model to craft systematic poisoning attacks aga

inst machine learning classifiers generating adversarial training examples, i.e. samples that look like genuine data points but that degrade the classifier's accuracy when used for training. We propose a Generative Adversarial Net with three components: generator, discriminator, and the target classifier. This approach allows us to model naturally the detectability constrains that can be expected in realistic attacks and to identify the regions of the underlying data distribution that can be more vulnerable to data poisoning. Our experimental evaluation shows the effectiveness of our attack to compromise machine learning classifiers, including deep networks.
****

Learnable Group Transform For Time-Series
Romain Cosentino,Behnaam Aazhang
We propose to undertake the problem of representation learning for time-series by considering a Group Transform approach. This framework allows us to, first, generalize classical time-frequency transformations such as the Wavelet Transform, and second, to enable the learnability of the representation. While the creation of the Wavelet Transform filter-bank relies on the sampling of the affine group in order to transform the mother filter, our approach allows for non-linear transformations of the mother filter by introducing the group of strictly increasing and continuous functions. The transformations induced by such a group enable us to span a larger class of signal representations. The sampling of this group can be optimized with respect to a specific loss and function and thus cast into a Deep Learning architecture. The experiments on diverse time-series datasets demonstrate the expressivity of this framework which competes with state-of-the-art performances.
****

From English to Foreign Languages: Transferring Pre-trained Language Models
Ke Tran
Pre-trained models have demonstrated their effectiveness in many downstream natural language processing (NLP) tasks. The availability of multilingual pre-trained models enables zero-shot transfer of NLP tasks from high resource languages to low resource ones. However, recent research in improving pre-trained models focuses heavily on English. While it is possible to train the latest neural architectures for other languages from scratch, it is undesirable due to the required amount of compute. In this work, we tackle the problem of transferring an existing pre-trained model from English to other languages under a limited computational budget. With a single GPU, our approach can obtain a foreign BERT-base model within a day and a foreign BERT-large within two days. Furthermore, evaluating our models on six languages, we demonstrate that our models are better than multilingual BERT on two zero-shot tasks: natural language inference and dependency parsing.
****

CoPhy: Counterfactual Learning of Physical Dynamics
Fabien Baradel,Natalia Neverova,Julien Mille,Greg Mori,Christian Wolf
Understanding causes and effects in mechanical systems is an essential component of reasoning in the physical world. This work poses a new problem of counterfactual learning of object mechanics from visual input. We develop the CoPhy benchmark to assess the capacity of the state-of-the-art models for causal physical reasoning in a synthetic 3D environment and propose a model for learning the physical dynamics in a counterfactual setting. Having observed a mechanical experiment that involves, for example, a falling tower of blocks, a set of bouncing balls or colliding objects, we learn to predict how its outcome is affected by an arbitrary intervention on its initial conditions, such as displacing one of the objects in the scene. The alternative future is predicted given the altered past and a latent representation of the confounders learned by the model in an end-to-end fashion with no supervision. We compare against feedforward video prediction baselines and show how observing alternative experiences allows the network to capture latent physical properties of the environment, which results in significantly more accurate predictions at the level of super human performance.
****

Semi-Supervised Few-Shot Learning with Prototypical Random Walks
Ahmed Ayyad,Nassir Navab,Mohamed Elhoseiny,Shadi Albarqouni
Learning from a few examples is a key characteristic of human intelligence that inspired machine learning researchers to build data-efficient AI models. Recent progress has shown that few-shot learning can be improved with access to unlabelled data, known as semi-supervised few-shot learning(SS-FSL). We introduce an SS-FSL approach, dubbed as Prototypical Random Walk Networks (PRWN), built on top of Prototypical Networks (PN). We develop a random walk semi-supervised loss that enables the network to learn representations that are compact and well-separated. Our work is related to the very recent development on graph-based approaches for few-shot learning. However, we show that achieved compact and well-separated class embeddings can be achieved by our prototypical random walk notion without needing additional graph-NN parameters or requiring a transductive setting where collective test set is provided. Our model outperforms prior art in most benchmarks with significant improvements in some cases. For example, in a mini-Imagenet 5-shot classification task, we obtain 69.65% accuracy to the 64.59% state-of-the-art. Our model, trained with 40% of the data as labelled, compares competitively against fully supervised prototypical networks, trained on 100% of the labels, even outperforming it in the 1-shot mini-Imagenet case with 50.89% to 49.4% accuracy. We also show that our model is resistant to distractors, unlabeled data that does not belong to any of the training classes, and hence reflecting robustness to labelled/unlabelled class distribution mismatch. We also performed a challenging discriminative power test, showing a relative improvement on top of the baseline of 14% on 20 classes on mini-Imagenet and 60% on 800 classes on Omniglot.
****
Why Convolutional Networks Learn Oriented Bandpass Filters: A Hypothesis
Richard P. Wildes
It has been repeatedly observed that convolutional architectures when applied to image understanding tasks learn oriented bandpass filters. A standard explanation
of this result is that these filters reflect the structure of the images that they have
been exposed to during training: Natural images typically are locally composed of oriented contours at various scales and oriented bandpass filters are matched to such structure. The present paper offers an alternative explanation based not on the structure of images, but rather on the structure of convolutional architectures.
In particular, complex exponentials are the eigenfunctions of convolution. These eigenfunctions are defined globally; however, convolutional architectures operate locally. To enforce locality, one can apply a windowing function to the eigenfunctions, which leads to oriented bandpass filters as the natural operators
to be learned with convolutional architectures. From a representational point of view, these filters allow for a local systematic way to characterize and operate on
an image or other signal.
****
Improving SAT Solver Heuristics with Graph Networks and Reinforcement Learning
Vitaly Kurin,Saad Godil,Shimon Whiteson,Bryan Catanzaro
We present GQSAT, a branching heuristic in a Boolean SAT solver trained with value-based reinforcement learning (RL) using Graph Neural Networks for function approximation. Solvers using GQSAT are complete SAT solvers that either provide a satisfying assignment or a proof of unsatisfiability, which is required for many SAT applications. The branching heuristic commonly used in SAT solvers today suffers from bad decisions during their warm-up period, whereas GQSAT has been trained to examine the structure of the particular problem instance to make better decisions at the beginning of the search. Training GQSAT is data efficient and does not require elaborate dataset preparation or feature engineering to train. We train GQSAT on small SAT problems using RL interfacing with an existing SAT so

lver. We show that GQSAT is able to reduce the number of iterations required to solve SAT problems by 2-3X, and it generalizes to unsatisfiable SAT instances, a s well as to problems with 5X more variables than it was trained on. We also sho w that, to a lesser extent, it generalizes to SAT problems from different domain s by evaluating it on graph coloring. Our experiments show that augmenting SAT s olvers with agents trained with RL and graph neural networks can improve perform ance on the SAT search problem.
****
Unsupervised Out-of-Distribution Detection with Batch Normalization
Jiaming Song,Yang Song,Stefano Ermon
Likelihood from a generative model is a natural statistic for detecting out-of-d istribution (OoD) samples. However, generative models have been shown to assign higher likelihood to OoD samples compared to ones from the training distribution , preventing simple threshold-based detection rules. We demonstrate that OoD det ection fails even when using more sophisticated statistics based on the likeliho ods of individual samples. To address these issues, we propose a new method that leverages batch normalization. We argue that batch normalization for generative models challenges the traditional \emph{i.i.d.} data assumption and changes the corresponding maximum likelihood objective. Based on this insight, we propose t o exploit in-batch dependencies for OoD detection. Empirical results suggest tha t this leads to more robust detection for high-dimensional images.
****
Understanding the Limitations of Variational Mutual Information Estimators
Jiaming Song,Stefano Ermon
Variational approaches based on neural networks are showing promise for estimati ng mutual information (MI) between high dimensional variables. However, they can be difficult to use in practice due to poorly understood bias/variance tradeoff s. We theoretically show that, under some conditions, estimators such as MINE ex hibit variance that could grow exponentially with the true amount of underlying MI. We also empirically demonstrate that existing estimators fail to satisfy ba sic self-consistency properties of MI, such as data processing and additivity un der independence. Based on a unified perspective of variational approaches, we d evelop a new estimator that focuses on variance reduction. Empirical results on standard benchmark tasks demonstrate that our proposed estimator exhibits improv ed bias-variance trade-offs on standard benchmark tasks.
****
Latent Question Reformulation and Information Accumulation for Multi-Hop Machine Reading
Quentin Grail,Julien Perez,Eric Gaussier
Multi-hop text-based question-answering is a current challenge in machine compre hension.
This task requires to sequentially integrate facts from multiple passages to ans wer complex natural language questions.
In this paper, we propose a novel architecture, called the Latent Question Refor mulation Network (LQR-net), a multi-hop and parallel attentive network designed for question-answering tasks that require reasoning capabilities.
LQR-net is composed of an association of \textbf{reading modules} and \textbf{re formulation modules}.
The purpose of the reading module is to produce a question-aware representation of the document.
From this document representation, the reformulation module extracts essential e lements to calculate an updated representation of the question.
This updated question is then passed to the following hop.
We evaluate our architecture on the \hotpotqa question-answering dataset designe d to assess multi-hop reasoning capabilities.
Our model achieves competitive results on the public leaderboard and outperforms the best current \textit{published} models in terms of Exact Match (EM) and $F_ 1$ score.
Finally, we show that an analysis of the sequential reformulations can provide i nterpretable reasoning paths.

****
Hamiltonian Generative Networks
Peter Toth,Danilo J. Rezende,Andrew Jaegle,Sébastien Racanière,Aleksandar Botev,Irina Higgins
The Hamiltonian formalism plays a central role in classical and quantum physics. Hamiltonians are the main tool for modelling the continuous time evolution of systems with conserved quantities, and they come equipped with many useful properties, like time reversibility and smooth interpolation in time. These properties are important for many machine learning problems - from sequence prediction to reinforcement learning and density modelling - but are not typically provided out of the box by standard tools such as recurrent neural networks. In this paper, we introduce the Hamiltonian Generative Network (HGN), the first approach capable of consistently learning Hamiltonian dynamics from high-dimensional observations (such as images) without restrictive domain assumptions. Once trained, we can use HGN to sample new trajectories, perform rollouts both forward and backward in time, and even speed up or slow down the learned dynamics. We demonstrate how a simple modification of the network architecture turns HGN into a powerful normalising flow model, called Neural Hamiltonian Flow (NHF), that uses Hamiltonian dynamics to model expressive densities. Hence, we hope that our work serves as a first practical demonstration of the value that the Hamiltonian formalism can bring to machine learning. More results and video evaluations are available at: http://tiny.cc/hgn
****
Customizing Sequence Generation with Multi-Task Dynamical Systems
Alex Bird,Christopher K. I. Williams
Dynamical system models (including RNNs) often lack the ability to adapt the sequence generation or prediction to a given context, limiting their real-world application. In this paper we show that hierarchical multi-task dynamical systems (MTDSs) provide direct user control over sequence generation, via use of a latent code z that specifies the customization to the
individual data sequence. This enables style transfer, interpolation and morphing within generated sequences. We show the MTDS can improve predictions via latent code interpolation, and avoid the long-term performance degradation of standard RNN approaches.
****
Feature Interaction Interpretability: A Case for Explaining Ad-Recommendation Systems via Neural Interaction Detection
Michael Tsang,Dehua Cheng,Hanpeng Liu,Xue Feng,Eric Zhou,Yan Liu
Recommendation is a prevalent application of machine learning that affects many users; therefore, it is important for recommender models to be accurate and interpretable. In this work, we propose a method to both interpret and augment the predictions of black-box recommender systems. In particular, we propose to interpret feature interactions from a source recommender model and explicitly encode these interactions in a target recommender model, where both source and target models are black-boxes. By not assuming the structure of the recommender system, our approach can be used in general settings. In our experiments, we focus on a prominent use of machine learning recommendation: ad-click prediction. We found that our interaction interpretations are both informative and predictive, e.g., significantly outperforming existing recommender models. What's more, the same approach to interpret interactions can provide new insights into domains even beyond recommendation, such as text and image classification.
****
Quantum Expectation-Maximization for Gaussian Mixture Models
Iordanis Kerenidis,Anupam Prakash,Alessandro Luongo
The Expectation-Maximization (EM) algorithm is a fundamental tool in unsupervised machine learning. It is often used as an efficient way to solve Maximum Likelihood (ML) and Maximum A Posteriori estimation problems, especially for models with latent variables. It is also the algorithm of choice to fit mixture models: generative models that represent unlabelled points originating from k different processes, as samples from k multivariate distributions. In this work we defi

ne and use a quantum version of EM to fit a Gaussian Mixture Model. Given quantu m access to a dataset of n vectors of dimension d, our algorithm has converg ence and precision guarantees similar to the classical algorithm, but the runtim e is only polylogarithmic in the number of elements in the training set, and is polynomial in other parameters - as the dimension of the feature space, and the number of components in the mixture. We generalize further the algorithm by fitt ing any mixture model of base distributions in the exponential family. We discus s the performance of the algorithm on datasets that are expected to be classifie d successfully by those algorithms, arguing that on those cases we can give stro ng guarantees on the runtime.

****

Behavior Regularized Offline Reinforcement Learning

Yifan Wu,George Tucker,Ofir Nachum

In reinforcement learning (RL) research, it is common to assume access to direct online interactions with the environment. However in many real-world applicatio ns, access to the environment is limited to a fixed offline dataset of logged ex perience. In such settings, standard RL algorithms have been shown to diverge or otherwise yield poor performance. Accordingly, much recent work has suggested a number of remedies to these issues. In this work, we introduce a general fram ework, behavior regularized actor critic (BRAC), to empirically evaluate recentl y proposed methods as well as a number of simple baselines across a variety of offline continuous control tasks. Surprisingly, we find that many of the technic al complexities introduced in recent methods are unnecessary to achieve strong p erformance. Additional ablations provide insights into which design choices matt er most in the offline RL setting.

****

Encoder-Agnostic Adaptation for Conditional Language Generation

Zachary M. Ziegler,Luke Melas-Kyriazi,Sebastian Gehrmann,Alexander M. Rush

Large pretrained language models have changed the way researchers approach discr iminative natural language understanding tasks, leading to the dominance of appr oaches that adapt a pretrained model for arbitrary downstream tasks. However, it is an open question how to use similar techniques for language generation. Earl y results in the encoder-agnostic setting have been mostly negative. In this wor k, we explore methods for adapting a pretrained language model to arbitrary cond itional input. We observe that pretrained transformer models are sensitive to la rge parameter changes during tuning. Therefore, we propose an adaptation that di rectly injects arbitrary conditioning into self attention, an approach we call p seudo self attention. Through experiments on four diverse conditional text gener ation tasks, we show that this encoder-agnostic technique outperforms strong bas elines, produces coherent generations, and is data-efficient.

****

Optimizing Data Usage via Differentiable Rewards

Xinyi Wang,Hieu Pham,Paul Michel,Antonios Anastasopoulos,Graham Neubig,Jaime Car bonell

To acquire a new skill, humans learn better and faster if a tutor, based on thei r current knowledge level, informs them of how much attention they should pay to particular content or practice problems. Similarly, a machine learning model co uld potentially be trained better with a scorer that “adapts” to its current lea rning state and estimates the importance of each training data instance. Trainin g such an adaptive scorer efficiently is a challenging problem; in order to prec isely quantify the effect of a data instance at a given time during the training , it is typically necessary to first complete the entire training process. To ef ficiently optimize data usage, we propose a reinforcement learning approach call ed Differentiable Data Selection (DDS). In DDS, we formulate a scorer network as a learnable function of the training data, which can be efficiently updated alo ng with the main model being trained. Specifically, DDS updates the scorer with an intuitive reward signal: it should up-weigh the data that has a similar gradi ent with a dev set upon which we would finally like to perform well. Without sig nificant computing overhead, DDS delivers strong and consistent improvements ove r several strong baselines on two very different tasks of machine translation an

d image classification.
****
Dropout: Explicit Forms and Capacity Control
Raman Arora,Peter L. Bartlett,Poorya Mianjy,Nathan Srebro
We investigate the capacity control provided by dropout in various machine learning problems. First, we study dropout for matrix sensing, where it induces a data-dependent regularizer that, in expectation, equals the weighted trace-norm of the product of the factors. In deep learning, we show that the data-dependent regularizer due to dropout directly controls the Rademacher complexity of the underlying class of deep neural networks. These developments enable us to give concrete generalization error bounds for the dropout algorithm in both matrix completion as well as training deep neural networks. We evaluate our theoretical findings on real-world datasets, including MovieLens, Fashion MNIST, and CIFAR-10.
****
Training Interpretable Convolutional Neural Networks towards Class-specific Filters
Haoyu Liang,Zhihao Ouyang,Hang Su,Yuyuan Zeng,Zihao He,Shu-tao Xia,Jun Zhu,Bo Zhang
Convolutional neural networks (CNNs) have often been treated as “black-box” and successfully used in a range of tasks. However, CNNs still suffer from the problem of filter ambiguity – an intricate many-to-many mapping relationship between filters and features, which undermines the models’ interpretability. To interpret CNNs, most existing works attempt to interpret a pre-trained model, while neglecting to reduce the filter ambiguity hidden behind. To this end, we propose a simple but effective strategy for training interpretable CNNs. Specifically, we propose a novel Label Sensitive Gate (LSG) structure to enable the model to learn disentangled filters in a supervised manner, in which redundant channels experience a periodical shutdown as flowing through a learnable gate varying with input labels. To reduce redundant filters during training, LSG is constrained with a sparsity regularization. In this way, such training strategy imposes each filter’s attention to just one or few classes, namely class-specific. Extensive experiments demonstrate the fabulous performance of our method in generating sparse and highly label- related representation of the input. Moreover, comparing to the standard training strategy, our model displays less redundancy and stronger interpretability.

****
Faster Neural Network Training with Data Echoing
Dami Choi,Alexandre Passos,Christopher J. Shallue,George E. Dahl
In the twilight of Moore's law, GPUs and other specialized hardware accelerators have dramatically sped up neural network training. However, earlier stages of the training pipeline, such as disk I/O and data preprocessing, do not run on accelerators. As accelerators continue to improve, these earlier stages will increasingly become the bottleneck. In this paper, we introduce “data echoing,” which reduces the total computation used by earlier pipeline stages and speeds up training whenever computation upstream from accelerators dominates the training time. Data echoing reuses (or “echoes”) intermediate outputs from earlier pipeline stages in order to reclaim idle capacity. We investigate the behavior of different data echoing algorithms on various workloads, for various amounts of echoing, and for various batch sizes. We find that in all settings, at least one data echoing algorithm can match the baseline's predictive performance using less upstream computation. We measured a factor of 3.25 decrease in wall-clock time for ResNet-50 on ImageNet when reading training data over a network.
****
Kronecker Attention Networks
Hongyang Gao,Zhengyang Wang,Shuiwang Ji
Attention operators have been applied on both 1-D data like texts and higher-order data such as images and videos. Use of attention operators on high-order data requires flattening of the spatial or spatial-temporal dimensions into a vector, which is assumed to follow a multivariate normal distribution. This not only i

ncurs excessive requirements on computational resources, but also fails to prese
rve structures in data. In this work, we propose to avoid flattening by developi
ng Kronecker attention operators (KAOs) that operate on high-order tensor data d
irectly. KAOs lead to dramatic reductions in computational resources. Moreover,
we analyze KAOs theoretically from a probabilistic perspective and point out tha
t KAOs assume the data follow matrix-variate normal distributions. Experimental
results show that KAOs reduce the amount of required computational resources by
a factor of hundreds, with larger factors for higher-dimensional and higher-orde
r data. Results also show that networks with KAOs outperform models without atte
ntion, while achieving competitive performance as those with original attention
operators.
****

Farkas layers: don't shift the data, fix the geometry
Aram-Alexandre Pooladian,Chris Finlay,Adam M Oberman
Successfully training deep neural networks often requires either {batch normaliz
ation}, appropriate {weight initialization}, both of which come with their own c
hallenges. We propose an alternative, geometrically motivated method for trainin
g. Using elementary results from linear programming, we introduce Farkas layers:
 a method that ensures at least one neuron is active at a given layer. Focusing
on residual networks with ReLU activation, we empirically demonstrate a signific
ant improvement in training capacity in the absence of batch normalization or me
thods of initialization across a broad range of network sizes on benchmark datas
ets.
****

Unsupervised Model Selection for Variational Disentangled Representation Learnin
g
Sunny Duan,Loic Matthey,Andre Saraiva,Nick Watters,Chris Burgess,Alexander Lerch
ner,Irina Higgins
Disentangled representations have recently been shown to improve fairness, data
efficiency and generalisation in simple supervised and reinforcement learning ta
sks. To extend the benefits of disentangled representations to more complex doma
ins and practical applications, it is important to enable hyperparameter tuning
and model selection of existing unsupervised approaches without requiring access
 to ground truth attribute labels, which are not available for most datasets. Th
is paper addresses this problem by introducing a simple yet robust and reliable
method for unsupervised disentangled model selection. We show that our approach
performs comparably to the existing supervised alternatives across 5400 models f
rom six state of the art unsupervised disentangled representation learning model
 classes. Furthermore, we show that the ranking produced by our approach correla
tes well with the final task performance on two different domains.
****

Sticking to the Facts: Confident Decoding for Faithful Data-to-Text Generation
Ran Tian,Shashi Narayan,Thibault Sellam,Ankur P. Parikh
Neural conditional text generation systems have achieved significant progress in
 recent years, showing the ability to produce highly fluent text. However, the i
nherent lack of controllability in these systems allows them to hallucinate fact
ually incorrect phrases that are unfaithful to the source, making them often uns
uitable for many real world systems that require high degrees of precision. In t
his work, we propose a novel confidence oriented decoder that assigns a confiden
ce score to each target position. This score is learned in training using a vari
ational Bayes objective, and can be leveraged at inference time using a calibrat
ion technique to promote more faithful generation. Experiments on a structured d
ata-to-text dataset -- WikiBio -- show that our approach is more faithful to the
 source than existing state-of-the-art approaches, according to both automatic m
etrics and human evaluation.
****

How much Position Information Do Convolutional Neural Networks Encode?
Md Amirul Islam*,Sen Jia*,Neil D. B. Bruce
In contrast to fully connected networks, Convolutional Neural Networks (CNNs) ac
hieve efficiency by learning weights associated with local filters with a finite

spatial extent. An implication of this is that a filter may know what it is looking at, but not where it is positioned in the image. Information concerning absolute position is inherently useful, and it is reasonable to assume that deep CNNs may implicitly learn to encode this information if there is a means to do so. In this paper, we test this hypothesis revealing the surprising degree of absolute position information that is encoded in commonly used neural networks. A comprehensive set of experiments show the validity of this hypothesis and shed light on how and where this information is represented while offering clues to where positional information is derived from in deep CNNs.
****

A Theoretical Analysis of the Number of Shots in Few-Shot Learning
Tianshi Cao,Marc T Law,Sanja Fidler
Few-shot classification is the task of predicting the category of an example from a set of few labeled examples. The number of labeled examples per category is called the number of shots (or shot number). Recent works tackle this task through meta-learning, where a meta-learner extracts information from observed tasks during meta-training to quickly adapt to new tasks during meta-testing. In this formulation, the number of shots exploited during meta-training has an impact on the recognition performance at meta-test time. Generally, the shot number used in meta-training should match the one used in meta-testing to obtain the best performance. We introduce a theoretical analysis of the impact of the shot number on Prototypical Networks, a state-of-the-art few-shot classification method. From our analysis, we propose a simple method that is robust to the choice of shot number used during meta-training, which is a crucial hyperparameter. The performance of our model trained for an arbitrary meta-training shot number shows great performance for different values of meta-testing shot numbers. We experimentally demonstrate our approach on different few-shot classification benchmarks.
****

Event extraction from unstructured Amharic text
Ephrem Tadesse,Rosa Tsegaye,Kuulaa Qaqqabaa
In information extraction, event extraction is one of the types that extract the specific knowledge of certain incidents from texts. Event extraction has been done on different languages texts but not on one of the Semitic language Amharic. In this study, we present a system that extracts an event from unstructured Amharic text. The system has designed by the integration of supervised machine learning and rule-based approaches together. We call it a hybrid system. The model from the supervised machine learning detects events from the text, then, handcrafted rules and the rule-based rules extract the event from the text. The hybrid system has compared with the standalone rule-based method that is well known for event extraction. The study has shown that the hybrid system has outperformed the standalone rule-based method. For the event extraction, we have been extracting event arguments. Event arguments identify event triggering words or phrases that clearly express the occurrence of the event. The event argument attributes can be verbs, nouns, occasionally adjectives such as ■■■/wedding and time as well.
****

Representation Learning for Remote Sensing: An Unsupervised Sensor Fusion Approach
Aidan M. Swope,Xander H. Rudelis,Kyle T. Story
In the application of machine learning to remote sensing, labeled data is often scarce or expensive, which impedes the training of powerful models like deep convolutional neural networks. Although unlabeled data is abundant, recent self-supervised learning approaches are ill-suited to the remote sensing domain. In addition, most remote sensing applications currently use only a small subset of the multi-sensor, multi-channel information available, motivating the need for fused multi-sensor representations. We propose a new self-supervised training objective, Contrastive Sensor Fusion, which exploits coterminous data from multiple sources to learn useful representations of every possible combination of those sources. This method uses information common across multiple sensors and bands by training a single model to produce a representation that remains similar when any

subset of its input channels is used. Using a dataset of 47 million unlabeled co terminous image triplets, we train an encoder to produce semantically meaningful representations from any possible combination of channels from the input sensor s. These representations outperform fully supervised ImageNet weights on a remot e sensing classification task and improve as more sensors are fused.
****
Dynamical Distance Learning for Semi-Supervised and Unsupervised Skill Discovery
Kristian Hartikainen,Xinyang Geng,Tuomas Haarnoja,Sergey Levine
Reinforcement learning requires manual specification of a reward function to lea rn a task. While in principle this reward function only needs to specify the tas k goal, in practice reinforcement learning can be very time-consuming or even in feasible unless the reward function is shaped so as to provide a smooth gradient towards a successful outcome. This shaping is difficult to specify by hand, par ticularly when the task is learned from raw observations, such as images. In thi s paper, we study how we can automatically learn dynamical distances: a measure of the expected number of time steps to reach a given goal state from any other state. These dynamical distances can be used to provide well-shaped reward funct ions for reaching new goals, making it possible to learn complex tasks efficient ly. We show that dynamical distances can be used in a semi-supervised regime, wh ere unsupervised interaction with the environment is used to learn the dynamical distances, while a small amount of preference supervision is used to determine the task goal, without any manually engineered reward function or goal examples. We evaluate our method both on a real-world robot and in simulation. We show th at our method can learn to turn a valve with a real-world 9-DoF hand, using raw image observations and just ten preference labels, without any other supervision . Videos of the learned skills can be found on the project website: https://site s.google.com/view/dynamical-distance-learning
****
Project and Forget: Solving Large Scale Metric Constrained Problems
Anna C. Gilbert,Rishi Sonthalia
Given a set of distances amongst points, determining what metric representation is most “consistent” with the input distances or the metric that captures the re levant geometric features of the data is a key step in many machine learning alg orithms. In this paper, we focus on metric constrained problems, a class of opti mization problems with metric constraints. In particular, we identify three type s of metric constrained problems: metric nearness Brickell et al. (2008), weight ed correlation clustering on general graphs Bansal et al. (2004), and metric lea rning Bellet et al. (2013); Davis et al. (2007). Because of the large number of constraints in these problems, however, researchers have been forced to restrict either the kinds of metrics learned or the size of the problem that can be solv ed.
We provide an algorithm, PROJECT AND FORGET, that uses Bregman projections with cutting planes, to solve metric constrained problems with many (possibly exponen tially) inequality constraints. We also prove that our algorithm converges to th e global optimal solution. Additionally, we show that the optimality error (L2 d istance of the current iterate to the optimal) asymptotically decays at an expon ential rate. We show that using our method we can solve large problem instances of three types of metric constrained problems, out-performing all state of the a rt methods with respect to CPU times and problem sizes.
****
On the Variance of the Adaptive Learning Rate and Beyond
Liyuan Liu,Haoming Jiang,Pengcheng He,Weizhu Chen,Xiaodong Liu,Jianfeng Gao,Jiaw ei Han
The learning rate warmup heuristic achieves remarkable success in stabilizing tr aining, accelerating convergence and improving generalization for adaptive stoch astic optimization algorithms like RMSprop and Adam. Pursuing the theory behind warmup, we identify a problem of the adaptive learning rate -- its variance is p roblematically large in the early stage, and presume warmup works as a variance reduction technique. We provide both empirical and theoretical evidence to verif y our hypothesis. We further propose Rectified Adam (RAdam), a novel variant of

Adam, by introducing a term to rectify the variance of the adaptive learning rat
e. Experimental results on image classification, language modeling, and neural m
achine translation verify our intuition and demonstrate the efficacy and robustn
ess of RAdam.
****
Translation Between Waves, wave2wave
Tsuyoshi Okita,Hirotaka Hachiya,Sozo Inoue,Naonori Ueda
The understanding of sensor data has been greatly improved by advanced deep lear
ning methods with big data. However, available sensor data in the real world are
 still limited, which is called the opportunistic sensor problem. This paper pr
oposes a new variant of neural machine translation seq2seq to deal with continuo
us signal waves by introducing the window-based (inverse-) representation to ada
ptively represent partial shapes of waves and the iterative back-translation mod
el for high-dimensional data. Experimental results are shown for two real-life
data: earthquake and activity translation. The performance improvements of one-
dimensional data was about 46 % in test loss and that of high-dimensional data w
as about 1625 % in perplexity with regard to the original seq2seq.

****
Quantifying the Cost of Reliable Photo Authentication via High-Performance Learn
ed Lossy Representations
Pawel Korus,Nasir Memon
Detection of photo manipulation relies on subtle statistical traces, notoriously
 removed by aggressive lossy compression employed online. We demonstrate that en
d-to-end modeling of complex photo dissemination channels allows for codec optim
ization with explicit provenance objectives. We design a lightweight trainable l
ossy image codec, that delivers competitive rate-distortion performance, on par
with best hand-engineered alternatives, but has lower computational footprint on
 modern GPU-enabled platforms. Our results show that significant improvements in
 manipulation detection accuracy are possible at fractional costs in bandwidth/s
torage. Our codec improved the accuracy from 37% to 86% even at very low bit-rat
es, well below the practicality of JPEG (QF 20).
****
Improving End-to-End Object Tracking Using Relational Reasoning
Fabian B. Fuchs,Adam R. Kosiorek,Li Sun,Oiwi Parker Jones,Ingmar Posner
Relational reasoning, the ability to model interactions and relations between ob
jects, is valuable for robust multi-object tracking and pivotal for trajectory p
rediction. In this paper, we propose MOHART, a class-agnostic, end-to-end multi-
object tracking and trajectory prediction algorithm, which explicitly accounts f
or permutation invariance in its relational reasoning. We explore a number of pe
rmutation invariant architectures and show that multi-headed self-attention outp
erforms the provided baselines and better accounts for complex physical interact
ions in a challenging toy experiment. We show on three real-world tracking datas
ets that adding relational reasoning capabilities in this way increases the trac
king and trajectory prediction performance, particularly in the presence of ego-
motion, occlusions, crowded scenes, and faulty sensor inputs. To the best of our
 knowledge, MOHART is the first fully end-to-end multi-object tracking from visi
on approach applied to real-world data reported in the literature.
****
Attention Privileged Reinforcement Learning for Domain Transfer
Sasha Salter,Dushyant Rao,Markus Wulfmeier,Raia Hadsell,Ingmar Posner
Applying reinforcement learning (RL) to physical systems presents notable challe
nges, given requirements regarding sample efficiency, safety, and physical const
raints compared to simulated environments. To enable transfer of policies traine
d in simulation, randomising simulation parameters leads to more robust policies
, but also in significantly extended training time. In this paper, we exploit ac
cess to privileged information (such as environment states) often available in s
imulation, in order to improve and accelerate learning over randomised environme
nts. We introduce Attention Privileged Reinforcement Learning (APRiL), which equ
ips the agent with an attention mechanism and makes use of state information in

simulation, learning to align attention between state- and image-based policies while additionally sharing generated data. During deployment we can apply the image-based policy to remove the requirement of access to additional information. We experimentally demonstrate accelerated and more robust learning on a number of diverse domains, leading to improved final performance for environments both within and outside the training distribution.
****

Sliced Cramer Synaptic Consolidation for Preserving Deeply Learned Representations

Soheil Kolouri,Nicholas A. Ketz,Andrea Soltoggio,Praveen K. Pilly

Deep neural networks suffer from the inability to preserve the learned data representation (i.e., catastrophic forgetting) in domains where the input data distribution is non-stationary, and it changes during training. Various selective synaptic plasticity approaches have been recently proposed to preserve network parameters, which are crucial for previously learned tasks while learning new tasks. We explore such selective synaptic plasticity approaches through a unifying lens of memory replay and show the close relationship between methods like Elastic Weight Consolidation (EWC) and Memory-Aware-Synapses (MAS). We then propose a fundamentally different class of preservation methods that aim at preserving the distribution of internal neural representations for previous tasks while learning a new one. We propose the sliced Cram\'{e}r distance as a suitable choice for such preservation and evaluate our Sliced Cramer Preservation (SCP) algorithm through extensive empirical investigations on various network architectures in both supervised and unsupervised learning settings. We show that SCP consistently utilizes the learning capacity of the network better than online-EWC and MAS methods on various incremental learning tasks.
****

On Variational Learning of Controllable Representations for Text without Supervision

Peng Xu,Yanshuai Cao,Jackie Chi Kit Cheung

The variational autoencoder (VAE) has found success in modelling the manifold of natural images on certain datasets, allowing meaningful images to be generated while interpolating or extrapolating in the latent code space, but it is unclear whether similar capabilities are feasible for text considering its discrete nature. In this work, we investigate the reason why unsupervised learning of controllable representations fails for text. We find that traditional sequence VAEs can learn disentangled representations through their latent codes to some extent, but they often fail to properly decode when the latent factor is being manipulated, because the manipulated codes often land in holes or vacant regions in the aggregated posterior latent space, which the decoding network is not trained to process. Both as a validation of the explanation and as a fix to the problem, we propose to constrain the posterior mean to a learned probability simplex, and performs manipulation within this simplex. Our proposed method mitigates the latent vacancy problem and achieves the first success in unsupervised learning of controllable representations for text. Empirically, our method significantly outperforms unsupervised baselines and is competitive with strong supervised approaches on text style transfer. Furthermore, when switching the latent factor (e.g., topic) during a long sentence generation, our proposed framework can often complete the sentence in a seemingly natural way -- a capability that has never been attempted by previous methods.
****

Disentangled Representation Learning with Sequential Residual Variational Autoencoder

Nanxiang Li,Shabnam Ghaffarzadegan,Liu Ren

Recent advancements in unsupervised disentangled representation learning focus on extending the variational autoencoder (VAE) with an augmented objective function to balance the trade-off between disentanglement and reconstruction. We propose Sequential Residual Variational Autoencoder (SR-VAE) that defines a "Residual learning" mechanism as the training regime instead of the augmented objective function. Our proposed solution deploys two important ideas in a single framework

: (1) learning from the residual between the input data and the accumulated reconstruction of sequentially added latent variables; (2) decomposing the reconstruction into decoder output and a residual term. This formulation encourages the disentanglement in the latent space by inducing explicit dependency structure, and reduces the bottleneck of VAE by adding the residual term to facilitate reconstruction. More importantly, SR-VAE eliminates the hyperparameter tuning, a crucial step for the prior state-of-the-art performance using the objective function augmentation approach. We demonstrate both qualitatively and quantitatively that SR-VAE improves the state-of-the-art unsupervised disentangled representation learning on a variety of complex datasets.

****

Improved Training Speed, Accuracy, and Data Utilization via Loss Function Optimization

Santiago Gonzalez,Risto Miikkulainen

As the complexity of neural network models has grown, it has become increasingly important to optimize their design automatically through metalearning. Methods for discovering hyperparameters, topologies, and learning rate schedules have lead to significant increases in performance. This paper shows that loss functions can be optimized with metalearning as well, and result in similar improvements. The method, Genetic Loss-function Optimization (GLO), discovers loss functions de novo, and optimizes them for a target task. Leveraging techniques from genetic programming, GLO builds loss functions hierarchically from a set of operators and leaf nodes. These functions are repeatedly recombined and mutated to find an optimal structure, and then a covariance-matrix adaptation evolutionary strategy (CMA-ES) is used to find optimal coefficients. Networks trained with GLO loss functions are found to outperform the standard cross-entropy loss on standard image classification tasks. Training with these new loss functions requires fewer steps, results in lower test error, and allows for smaller datasets to be used. Loss function optimization thus provides a new dimension of metalearning, and constitutes an important step towards AutoML.

****

Using Hindsight to Anchor Past Knowledge in Continual Learning

Arslan Chaudhry,Albert Gordo,David Lopez-Paz,Puneet K. Dokania,Philip Torr

In continual learning, the learner faces a stream of data whose distribution changes over time. Modern neural networks are known to suffer under this setting, as they quickly forget previously acquired knowledge. To address such catastrophic forgetting, state-of-the-art continual learning methods implement different types of experience replay, re-learning on past data stored in a small buffer known as episodic memory. In this work, we complement experience replay with a meta-learning technique that we call anchoring: the learner updates its knowledge on the current task, while keeping predictions on some anchor points of past tasks intact. These anchor points are learned using gradient-based optimization as to maximize forgetting of the current task, in hindsight, when the learner is fine-tuned on the episodic memory of past tasks. Experiments on several supervised learning benchmarks for continual learning demonstrate that our approach improves the state of the art in terms of both accuracy and forgetting metrics and for various sizes of episodic memories.

****

Empirical confidence estimates for classification by deep neural networks

Chris Finlay,Adam M. Oberman

How well can we estimate the probability that the classification predicted by a deep neural network is correct (or in the Top 5)? It is well-known that the softmax values of the network are not estimates of the probabilities of class labels. However, there is a misconception that these values are not informative. We define the notion of implied loss and prove that if an uncertainty measure is an implied loss, then low uncertainty means high probability of correct (or Top-k) classification on the test set. We demonstrate empirically that these values can be used to measure the confidence that the classification is correct. Our method is simple to use on existing networks: we proposed confidence measures for Top-k which can be evaluated by binning values on the test set.

****
Learning Numeral Embedding
Chengyue Jiang,Zhonglin Nian,Kaihao Guo,Shanbo Chu,Yinggong Zhao,Libin Shen,Kewei Tu
Word embedding is an essential building block for deep learning methods for natural language processing. Although word embedding has been extensively studied over the years, the problem of how to effectively embed numerals, a special subset of words, is still underexplored. Existing word embedding methods do not learn numeral embeddings well because there are an infinite number of numerals and their individual appearances in training corpora are highly scarce.
In this paper, we propose two novel numeral embedding methods that can handle the out-of-vocabulary (OOV) problem for numerals. We first induce a finite set of prototype numerals using either a self-organizing map or a Gaussian mixture model. We then represent the embedding of a numeral as a weighted average of the prototype number embeddings. Numeral embeddings represented in this manner can be plugged into existing word embedding learning approaches such as skip-gram for training.
We evaluated our methods and showed its effectiveness on four intrinsic and extrinsic tasks: word similarity, embedding numeracy, numeral prediction, and sequence labeling.
****
Localized Generations with Deep Neural Networks for Multi-Scale Structured Datasets
Yoshihiro Nagano,Shiro Takagi,Yuki Yoshida,Masato Okada
Extracting the hidden structure of the external environment is an essential component of intelligent agents and human learning. The real-world datasets that we are interested in are often characterized by the locality: the change in the structural relationship between the data points depending on location in observation space. The local learning approach extracts semantic representations for these datasets by training the embedding model from scratch for each local neighborhood, respectively. However, this approach is only limited to use with a simple model, since the complex model, including deep neural networks, requires a massive amount of data and extended training time. In this study, we overcome this trade-off based on the insight that the real-world dataset often shares some structural similarity between each neighborhood. We propose to utilize the embedding model for the other local structure as a weak form of supervision. Our proposed model, the Local VAE, generalize the Variational Autoencoder to have the different model parameters for each local subset and train these local parameters by the gradient-based meta-learning. Our experimental results showed that the Local VAE succeeded in learning the semantic representations for the dataset with local structure, including the 3D Shapes Dataset, and generated high-quality images.
****
AlgoNet: C^∞ Smooth Algorithmic Neural Networks
Felix Petersen,Christian Borgelt,Oliver Deussen
Artificial neural networks have revolutionized many areas of computer science in recent years, providing solutions to a number of previously unsolved problems. On the other hand, for many problems, classic algorithms exist, which typically exceed the accuracy and stability of neural networks.
To combine these two concepts, we present a new kind of neural networks—algorithmic neural networks (AlgoNets).
These networks integrate smooth versions of classic algorithms into the topology of neural networks.
A forward AlgoNet includes algorithmic layers into existing architectures to enhance performance and explainability while a backward AlgoNet enables solving inverse problems without or with only weak supervision.
In addition, we present the algonet package, a PyTorch based library that includes, inter alia, a smoothly evaluated programming language, a smooth 3D mesh renderer, and smooth sorting algorithms.
****
Temporal-difference learning for nonlinear value function approximation in the l

azy training regime
Andrea Agazzi,Jianfeng Lu
We discuss the approximation of the value function for infinite-horizon discount
ed Markov Reward Processes (MRP) with nonlinear functions trained with the Tempo
ral-Difference (TD) learning algorithm. We consider this problem under a certain
 scaling of the approximating function, leading to a regime called lazy training
. In this regime the parameters of the model vary only slightly during the learn
ing process, a feature that has recently been observed in the training of neural
 networks, where the scaling we study arises naturally, implicit in the initiali
zation of their parameters. Both in the under- and over-parametrized frameworks
, we prove exponential convergence to local, respectively global minimizers of t
he above algorithm in the lazy training regime. We then give examples of such co
nvergence results in the case of models that diverge if trained with non-lazy TD
 learning, and in the case of neural networks.
****
A Bayes-Optimal View on Adversarial Examples
Eitan Richardson,Yair Weiss
Adversarial attacks on CNN classifiers can make an imperceptible change to an in
put image and alter the classification result. The source of these failures is s
till poorly understood, and many explanations invoke the "unreasonably linear ex
trapolation" used by CNNs along with the geometry of high dimensions.
In this paper we show that similar attacks can be used against the Bayes-Optimal
 classifier for certain class distributions, while for others the optimal classi
fier is robust to such attacks. We present analytical results showing conditions
 on the data distribution under which all points can be made arbitrarily close t
o the optimal decision boundary and show that this can happen even when the clas
ses are easy to separate, when the ideal classifier has a smooth decision surfac
e and when the data lies in low dimensions. We introduce new datasets of realist
ic images of faces and digits where the Bayes-Optimal classifier can be calculat
ed efficiently and show that for some of these datasets the optimal classifier i
s robust and for others it is vulnerable to adversarial examples. In systematic
experiments with many such datasets, we find that standard CNN training consiste
ntly finds a vulnerable classifier even when the optimal classifier is robust wh
ile large-margin methods often find a robust classifier with the exact same trai
ning data. Our results suggest that adversarial vulnerability is not an unavoida
ble consequence of machine learning in high dimensions, and may often be a resul
t of suboptimal training methods used in current practice.
****
Efficient Content-Based Sparse Attention with Routing Transformers
Aurko Roy*,Mohammad Taghi Saffar*,David Grangier,Ashish Vaswani
Self-attention has recently been adopted for a wide range of sequence modeling
problems. Despite its effectiveness, self-attention suffers quadratic compute an
d
memory requirements with respect to sequence length. Successful approaches to
reduce this complexity focused on attention to local sliding windows or a small
set of locations independent of content. Our work proposes to learn dynamic
sparse attention patterns that avoid allocating computation and memory to attend
to content unrelated to the query of interest. This work builds upon two lines o
f
research: it combines the modeling flexibility of prior work on content-based sp
arse
attention with the efficiency gains from approaches based on local, temporal spa
rse
attention. Our model, the Routing Transformer, endows self-attention with a spar
se
routing module based on online k-means while reducing the overall complexity of
attention to O(n^{1.5}d) from O(n^2d) for sequence length n and hidden dimension
d. We show that our model outperforms comparable sparse attention models on
language modeling on Wikitext-103 (15.8 vs 18.3 perplexity) as well as on
image generation on ImageNet-64 (3.43 vs 3.44 bits/dim) while using fewer self-a

ttention layers.
Code will be open-sourced on acceptance.
****

Good Semi-supervised VAE Requires Tighter Evidence Lower Bound
Haozhe Feng,Kezhi Kong,Tianye Zhang,Siyue Xue,Wei Chen
Semi-supervised learning approaches based on generative models have now encounte red 3 challenges: (1) The two-stage training strategy is not robust. (2) Good se mi-supervised learning results and good generative performance can not be obtain ed at the same time. (3) Even at the expense of sacrificing generative performan ce, the semi-supervised classification results are still not satisfactory. To ad dress these problems, we propose One-stage Semi-suPervised Optimal Transport VAE (OSPOT-VAE), a one-stage deep generative model that theoretically unifies the g eneration and classification loss in one ELBO framework and achieves a tighter E LBO by applying the optimal transport scheme to the distribution of latent varia bles. We show that with tighter ELBO, our OSPOT-VAE surpasses the best semi-supe rvised generative models by a large margin across many benchmark datasets. For e xample, we reduce the error rate from 14.41% to 6.11% on Cifar-10 with 4k labels and achieve state-of-the-art performance with 25.30% on Cifar-100 with 10k labe ls. We also demonstrate that good generative models and semi-supervised results can be achieved simultaneously by OSPOT-VAE.
****

Option Discovery using Deep Skill Chaining
Akhil Bagaria,George Konidaris
Autonomously discovering temporally extended actions, or skills, is a longstandi ng goal of hierarchical reinforcement learning. We propose a new algorithm that combines skill chaining with deep neural networks to autonomously discover skill s in high-dimensional, continuous domains. The resulting algorithm, deep skill c haining, constructs skills with the property that executing one enables the agen t to execute another. We demonstrate that deep skill chaining significantly outp erforms both non-hierarchical agents and other state-of-the-art skill discovery techniques in challenging continuous control tasks.
****

HOPPITY: LEARNING GRAPH TRANSFORMATIONS TO DETECT AND FIX BUGS IN PROGRAMS
Elizabeth Dinella,Hanjun Dai,Ziyang Li,Mayur Naik,Le Song,Ke Wang
We present a learning-based approach to detect and fix a broad range of bugs in Javascript programs. We frame the problem in terms of learning a sequence of gra ph transformations: given a buggy program modeled by a graph structure, our mode l makes a sequence of predictions including the position of bug nodes and corres ponding graph edits to produce a fix. Unlike previous works that use deep neural networks, our approach targets bugs that are more complex and semantic in natur e (i.e.~bugs that require adding or deleting statements to fix). We have realize d our approach in a tool called HOPPITY. By training on 290,715 Javascript code change commits on Github, HOPPITY correctly detects and fixes bugs in 9,490 out of 36,361 programs in an end-to-end fashion. Given the bug location and type of the fix, HOPPITY also outperforms the baseline approach by a wide margin.
****

PowerSGD: Powered Stochastic Gradient Descent Methods for Accelerated Non-Convex Optimization
Jun Liu,Beitong Zhou,Weigao Sun,Ruijuan Chen,Claire J. Tomlin,Ye Yuan
In this paper, we propose a novel technique for improving the stochastic gradien t descent (SGD) method to train deep networks, which we term \emph{PowerSGD}. Th e proposed PowerSGD method simply raises the stochastic gradient to a certain po wer $\gamma\in[0,1]$ during iterations and introduces only one additional parame ter, namely, the power exponent γ (when $\gamma=1$, PowerSGD reduces to S GD). We further propose PowerSGD with momentum, which we term \emph{PowerSGDM}, and provide convergence rate analysis on both PowerSGD and PowerSGDM methods. Ex periments are conducted on popular deep learning models and benchmark datasets. Empirical results show that the proposed PowerSGD and PowerSGDM obtain faster in itial training speed than adaptive gradient methods, comparable generalization ability with SGD, and improved robustness to hyper-parameter selection and vanis

hing gradients. PowerSGD is essentially a gradient modifier via a nonlinear tran
sformation. As such, it is orthogonal and complementary to other techniques for
accelerating gradient-based optimization.
****

Deep Randomized Least Squares Value Iteration
Guy Adam,Tom Zahavy,Oron Anschel,Nahum Shimkin
Exploration while learning representations is one of the main challenges Deep
Reinforcement Learning (DRL) faces today. As the learned representation is depen
dant in the observed data, the exploration strategy has a crucial role. The popu
lar DQN algorithm has improved significantly the capabilities of Reinforcement
Learning (RL) algorithms to learn state representations from raw data, yet, it u
ses
a naive exploration strategy which is statistically inefficient. The Randomized
Least Squares Value Iteration (RLSVI) algorithm (Osband et al., 2016), on the
other hand, explores and generalizes efficiently via linearly parameterized valu
e
functions. However, it is based on hand-designed state representation that requi
res
prior engineering work for every environment. In this paper, we propose a Deep
Learning adaptation for RLSVI. Rather than using hand-design state representatio
n, we use a state representation that is being learned directly from the data by
 a
DQN agent. As the representation is being optimized during the learning process,
a key component for the suggested method is a likelihood matching mechanism,
which adapts to the changing representations. We demonstrate the importance of
the various properties of our algorithm on a toy problem and show that our metho
d
outperforms DQN in five Atari benchmarks, reaching competitive results with the
Rainbow algorithm.
****

Self-Supervised Policy Adaptation
Christopher Mutschler,Sebastian Pokutta
We consider the problem of adapting an existing policy when the environment repr
esentation changes. Upon a change of the encoding of the observations the agent
can no longer make use of its policy as it cannot correctly interpret the new ob
servations. This paper proposes Greedy State Representation Learning (GSRL) to t
ransfer the original policy by translating the environment representation back i
nto its original encoding. To achieve this GSRL samples observations from both t
he environment and a dynamics model trained from prior experience. This generate
s pairs of state encodings, i.e., a new representation from the environment and
a (biased) old representation from the forward model, that allow us to bootstrap
 a neural network model for state translation. Although early translations are u
nsatisfactory (as expected), the agent eventually learns a valid translation as
it minimizes the error between expected and observed environment dynamics. Our e
xperiments show the efficiency of our approach and that it translates the policy
 in considerably less steps than it would take to retrain the policy.
****

RTC-VAE: HARNESSING THE PECULIARITY OF TOTAL CORRELATION IN LEARNING DISENTANGL
ED REPRESENTATIONS
Ze Cheng,Juncheng B Li,Chenxu Wang,Jixuan Gu,Hao Xu,Xinjian Li,Florian Metze
In the problem of unsupervised learning of disentangled representations, one of
the promising methods is to penalize the total correlation of sampled latent var
i-ables. Unfortunately, this well-motivated strategy often fail to achieve dise
ntanglement due to a problematic difference between the sampled latent represent
ation and its corresponding mean representation. We provide a theoretical expla
nation that low total correlation of sample distribution cannot guarantee low to
tal correlation of the mean representation. We prove that for the mean represen
tation of arbitrarily high total correlation, there exist distributions of laten
t variables of abounded total correlation. However, we still believe that total
 correlation could be a key to the disentanglement of unsupervised representativ

e learning, and we propose a remedy, RTC-VAE, which rectifies the total correla tion penalty. Experiments show that our model has a more reasonable distributi on of the mean representation compared with baseline models, e.g.,β-TCVAE and Fa ctorVAE.
****

OmniNet: A unified architecture for multi-modal multi-task learning
Subhojeet Pramanik,Priyanka Agrawal,Aman Hussain
Transformer is a popularly used neural network architecture, especially for lang uage understanding. We introduce an extended and unified architecture that can b e used for tasks involving a variety of modalities like image, text, videos, etc . We propose a spatio-temporal cache mechanism that enables learning spatial dim ension of the input in addition to the hidden states corresponding to the tempor al input sequence. The proposed architecture further enables a single model to s upport tasks with multiple input modalities as well as asynchronous multi-task l earning, thus we refer to it as OmniNet. For example, a single instance of OmniN et can concurrently learn to perform the tasks of part-of-speech tagging, image captioning, visual question answering and video activity recognition. We demonst rate that training these four tasks together results in about three times compre ssed model while retaining the performance in comparison to training them indivi dually. We also show that using this neural network pre-trained on some modaliti es assists in learning unseen tasks such as video captioning and video question answering. This illustrates the generalization capacity of the self-attention me chanism on the spatio-temporal cache present in OmniNet.
****

Unified Probabilistic Deep Continual Learning through Generative Replay and Open Set Recognition
Martin Mundt,Sagnik Majumder,Iuliia Pliushch,Visvanathan Ramesh
We introduce a unified probabilistic approach for deep continual learning based on variational Bayesian inference with open set recognition. Our model combines a joint probabilistic encoder with a generative model and a linear classifier th at get shared across tasks. The open set recognition bounds the approximate post erior by fitting regions of high density on the basis of correctly classified da ta points and balances open set detection with recognition errors. Catastrophic forgetting is significantly alleviated through generative replay, where the open set recognition is used to sample from high density areas of the class specific posterior and reject statistical outliers. Our approach naturally allows for fo rward and backward transfer while maintaining past knowledge without the necessi ty of storing old data, regularization or inferring task labels. We demonstrate compelling results in the challenging scenario of incrementally expanding the si ngle-head classifier for both class incremental visual and audio classification tasks, as well as incremental learning of datasets across modalities.
****

TED: A Pretrained Unsupervised Summarization Model with Theme Modeling and Denoi sing
Ziyi Yang,Chenguang Zhu,Michael Zeng,Xuedong Huang,Eric Darve
Text summarization aims to extract essential information from a piece of text an d transform it into a concise version. Existing unsupervised abstractive summari zation models use recurrent neural networks framework and ignore abundant unlabe led corpora resources. In order to address these issues, we propose TED, a trans former-based unsupervised summarization system with dataset-agnostic pretraining . We first leverage the lead bias in news articles to pretrain the model on larg e-scale corpora. Then, we finetune TED on target domains through theme modeling and a denoising autoencoder to enhance the quality of summaries. Notably, TED ou tperforms all unsupervised abstractive baselines on NYT, CNN/DM and English Giga word datasets with various document styles. Further analysis shows that the summ aries generated by TED are abstractive and containing even higher proportions of novel tokens than those from supervised models.
****

V4D: 4D Convolutional Neural Networks for Video-level Representation Learning
Shiwen Zhang,Sheng Guo,Weilin Huang,Matthew R. Scott,Limin Wang

Most existing 3D CNN structures for video representation learning are clip-based methods, and do not consider video-level temporal evolution of spatio-temporal features. In this paper, we propose Video-level 4D Convolutional Neural Networks, namely V4D, to model the evolution of long-range spatio-temporal representation with 4D convolutions, as well as preserving 3D spatio-temporal representations with residual connections. We further introduce the training and inference methods for the proposed V4D. Extensive experiments are conducted on three video recognition benchmarks, where V4D achieves excellent results, surpassing recent 3D CNNs by a large margin.
****

ODE Analysis of Stochastic Gradient Methods with Optimism and Anchoring for Minimax Problems and GANs
Ernest K. Ryu,Kun Yuan,Wotao Yin
Despite remarkable empirical success, the training dynamics of generative adversarial networks (GAN), which involves solving a minimax game using stochastic gradients, is still poorly understood. In this work, we analyze last-iterate convergence of simultaneous gradient descent (simGD) and its variants under the assumption of convex-concavity, guided by a continuous-time analysis with differential equations. First, we show that simGD, as is, converges with stochastic sub-gradients under strict convexity in the primal variable. Second, we generalize optimistic simGD to accommodate an optimism rate separate from the learning rate and show its convergence with full gradients. Finally, we present anchored simGD, a new method, and show convergence with stochastic subgradients.
****

Learning to Represent Programs with Property Signatures
Augustus Odena,Charles Sutton
We introduce the notion of property signatures, a representation for programs and
program specifications meant for consumption by machine learning algorithms.
Given a function with input type τ_in and output type τ_out, a property is a function
of type: (τ_in, τ_out) → Bool that (informally) describes some simple property
of the function under consideration. For instance, if τ_in and τ_out are both lists
of the same type, one property might ask ‘is the input list the same length as the
output list?’. If we have a list of such properties, we can evaluate them all for our
function to get a list of outputs that we will call the property signature. Crucially,
we can ‘guess’ the property signature for a function given only a set of input/output
pairs meant to specify that function. We discuss several potential applications of
property signatures and show experimentally that they can be used to improve
over a baseline synthesizer so that it emits twice as many programs in less than
one-tenth of the time.
****

Unified recurrent network for many feature types
Alexander Stec,Diego Klabjan,Jean Utke
There are time series that are amenable to recurrent neural network (RNN) solutions when treated as sequences, but some series, e.g. asynchronous time series, provide a richer variation of feature types than current RNN cells take into account. In order to address such situations, we introduce a unified RNN that handles five different feature types, each in a different manner. Our RNN framework separates sequential features into two groups dependent on their frequency, which we call sparse and dense features, and which affect cell updates differently. Further, we also incorporate time features at the sequential level that relate to the time between specified events in the sequence and are used to modify the cell's memory state. We also include two types of static (whole sequence level) fea

tures, one related to time and one not, which are combined with the encoder outp
ut. The experiments show that the proposed modeling framework does increase perf
ormance compared to standard cells.
****

Improving Dirichlet Prior Network for Out-of-Distribution Example Detection
Jay Nandy

Determining the source of uncertainties in the predictions of AI systems are imp
ortant. It allows the users to act in an informative manner to improve the safet
y of such systems, applied to the real-world sensitive applications. Predictive
uncertainties can originate from the uncertainty in model parameters, data uncer
tainty or due to distributional mismatch between training and test examples. Whi
le recently, significant progress has been made to improve the predictive uncert
ainty estimation of deep learning models, most of these approaches either confla
te the distributional uncertainty with model uncertainty or data uncertainty. In
 contrast, the Dirichlet Prior Network (DPN) can model distributional uncertain
ty distinctly by parameterizing a prior Dirichlet over the predictive categorica
l distributions. However, their complex loss function by explicitly incorporatin
g KL divergence between Dirichlet distributions often makes the error surface
ill-suited to optimize for challenging datasets with multiple classes. In this p
aper, we present an improved DPN framework by proposing a novel loss function us
ing the standard cross-entropy loss along with a regularization term to control
the sharpness of the output Dirichlet distributions from the network. Our propos
ed loss function aims to improve the training efficiency of the DPN framework fo
r challenging classification tasks with large number of classes. In our experime
nts using synthetic and real datasets, we demonstrate that our DPN models can di
stinguish the distributional uncertainty from other uncertainty types. Our propo
sed approach significantly improves DPN frameworks and outperform the existing O
OD detectors on CIFAR-10 and CIFAR-100 dataset while also being able to recogniz
e distributional uncertainty distinctly.
****

Variational Autoencoders for Opponent Modeling in Multi-Agent Systems
Georgios Papoudakis,Stefano V. Albrecht

Multi-agent systems exhibit complex behaviors that emanate from the interactions
 of multiple agents in a shared environment. In this work, we are interested in
controlling one agent in a multi-agent system and successfully learn to interact
 with the other agents that have fixed policies. Modeling the behavior of other
agents (opponents) is essential in understanding the interactions of the agents
in the system. By taking advantage of recent advances in unsupervised learning,
we propose modeling opponents using variational autoencoders. Additionally, many
 existing methods in the literature assume that the opponent models have access
to opponent's observations and actions during both training and execution. To el
iminate this assumption, we propose a modification that attempts to identify the
 underlying opponent model, using only local information of our agent, such as i
ts observations, actions, and rewards. The experiments indicate that our opponen
t modeling methods achieve equal or greater episodic returns in reinforcement le
arning tasks against another modeling method.
****

Prototype Recalls for Continual Learning
Mengmi Zhang,Tao Wang,Joo Hwee Lim,Jiashi Feng

Continual learning is a critical ability of continually acquiring and transferri
ng knowledge without catastrophically forgetting previously learned knowledge. H
owever, enabling continual learning for AI remains a long-standing challenge. In
 this work, we propose a novel method, Prototype Recalls, that efficiently embed
s and recalls previously learnt knowledge to tackle catastrophic forgetting issu
e. In particular, we consider continual learning in classification tasks. For ea
ch classification task, our method learns a metric space containing a set of pro
totypes where embedding of the samples from the same class cluster around protot
ypes and class-representative prototypes are separated apart. To alleviate catas
trophic forgetting, our method preserves the embedding function from the samples
 to the previous metric space, through our proposed prototype recalls from previ

ous tasks. Specifically, the recalling process is implemented by replaying a sma
ll number of samples from previous tasks and correspondingly matching their embe
dding to their nearest class-representative prototypes. Compared with recent con
tinual learning methods, our contributions are fourfold: first, our method achie
ves the best memory retention capability while adapting quickly to new tasks. Se
cond, our method uses metric learning for classification and does not require ad
ding in new neurons given new object classes. Third, our method is more memory e
fficient since only class-representative prototypes need to be recalled. Fourth,
 our method suggests a promising solution for few-shot continual learning. Witho
ut tampering with the performance on initial tasks, our method learns novel conc
epts given a few training examples of each class in new tasks.
****

Generative Ratio Matching Networks
Akash Srivastava,Kai Xu,Michael U. Gutmann,Charles Sutton
Deep generative models can learn to generate realistic-looking images, but many
of the most effective methods are adversarial and involve a saddlepoint optimiza
tion, which requires a careful balancing of training between a generator network
 and a critic network. Maximum mean discrepancy networks (MMD-nets) avoid this i
ssue by using kernel as a fixed adversary, but unfortunately, they have not on t
heir own been able to match the generative quality of adversarial training. In t
his work, we take their insight of using kernels as fixed adversaries further an
d present a novel method for training deep generative models that does not invol
ve saddlepoint optimization. We call our method generative ratio matching or GRA
M for short. In GRAM, the generator and the critic networks do not play a zero-s
um game against each other, instead, they do so against a fixed kernel. Thus GRA
M networks are not only stable to train like MMD-nets but they also match and be
at the generative quality of adversarially trained generative networks.
****

Emergence of Compositional Language with Deep Generational Transmission
Michael Cogswell,Jiasen Lu,Stefan Lee,Devi Parikh,Dhruv Batra
Recent work has studied the emergence of language among deep reinforcement learn
ing agents that must collaborate to solve a task. Of particular interest are the
 factors that cause language to be compositional---i.e., express meaning by comb
ining words which themselves have meaning. Evolutionary linguists have found tha
t in addition to structural priors like those already studied in deep learning,
the dynamics of transmitting language from generation to generation contribute s
ignificantly to the emergence of compositionality. In this paper, we introduce
these cultural evolutionary dynamics into language emergence by periodically rep
lacing agents in a population to create a knowledge gap, implicitly inducing cul
tural transmission of language. We show that this implicit cultural transmission
 encourages the resulting languages to exhibit better compositional generalizati
on.
****

Deep Gradient Boosting -- Layer-wise Input Normalization of Neural Networks
Erhan Bilal
Stochastic gradient descent (SGD) has been the dominant optimization method for
training deep neural networks due to its many desirable properties. One of the m
ore remarkable and least understood quality of SGD is that it generalizes relati
vely well
on unseen data even when the neural network has millions of parameters. We hypot
hesize that in certain cases it is desirable to relax its intrinsic generalizati
on properties and introduce an extension of SGD called deep gradient boosting (D
GB). The key idea of DGB is that back-propagated gradients inferred using the ch
ain rule can be viewed as pseudo-residual targets of a gradient boosting problem
. Thus at each layer of a neural network the weight update is calculated by solv
ing the corresponding boosting problem using a linear base learner. The resultin
g weight update formula can also be viewed as a normalization procedure of the d
ata that arrives at each layer during the forward pass. When implemented as a se
parate input normalization layer (INN) the new architecture shows improved perfo
rmance on image recognition tasks when compared to the same architecture without

normalization layers. As opposed to batch normalization (BN), INN has no learnable parameters however it matches its performance on CIFAR10 and ImageNet classification tasks.
****
A Generalized Framework of Sequence Generation with Application to Undirected Sequence Models
Elman Mansimov,Alex Wang,Kyunghyun Cho
Undirected neural sequence models such as BERT (Devlin et al., 2019) have received renewed interest due to their success on discriminative natural language understanding tasks such as question-answering and natural language inference.
The problem of generating sequences directly from these models has received relatively little attention, in part because generating from such models departs significantly from the conventional approach of monotonic generation in directed sequence models. We investigate this problem by first proposing a generalized model of sequence generation that unifies decoding in directed and undirected models. The proposed framework models the process of generation rather than a resulting sequence, and under this framework, we derive various neural sequence models as special cases, such as autoregressive, semi-autoregressive, and refinement-based non-autoregressive models. This unification enables us to adapt decoding algorithms originally developed for directed sequence models to undirected models. We demonstrate this by evaluating various decoding strategies for a cross-lingual masked translation model (Lample and Conneau, 2019). Our experiments show that generation from undirected sequence models, under our framework, is competitive with the state of the art on WMT'14 English-German translation. We also demonstrate that the proposed approach enables constant-time translation with similar performance to linear-time translation from the same model by rescoring hypotheses with an autoregressive model.
****
In Search for a SAT-friendly Binarized Neural Network Architecture
Nina Narodytska,Hongce Zhang,Aarti Gupta,Toby Walsh
Analyzing the behavior of neural networks is one of the most pressing challenges in deep learning. Binarized Neural Networks are an important class of networks that allow equivalent representation in Boolean logic and can be analyzed formally with logic-based reasoning tools like SAT solvers. Such tools can be used to answer existential and probabilistic queries about the network, perform explanation generation, etc. However, the main bottleneck for all methods is their ability to reason about large BNNs efficiently. In this work, we analyze architectural design choices of BNNs and discuss how they affect the performance of logic-based reasoners. We propose changes to the BNN architecture and the training procedure to get a simpler network for SAT solvers without sacrificing accuracy on the primary task. Our experimental results demonstrate that our approach scales to larger deep neural networks compared to existing work for existential and probabilistic queries, leading to significant speed ups on all tested datasets.

****
Neural networks are a priori biased towards Boolean functions with low entropy
Chris Mingard,Joar Skalse,Guillermo Valle-Pérez,David Martínez-Rubio,Vladimir Mikulik,Ard A. Louis
Understanding the inductive bias of neural networks is critical to explaining their ability to generalise. Here,
for one of the simplest neural networks -- a single-layer perceptron with n input neurons, one output neuron, and no threshold bias term -- we prove that upon random initialisation of weights, the a priori probability $P(t)$ that it represents a Boolean function that classifies t points in $\{0,1\}^n$ as 1 has a remarkably simple form: $
P(t) = 2^{-n} \,\, {\rm for} \,\, 0\leq t < 2^n$.
Since a perceptron can express far fewer Boolean functions with small or large values of t (low "entropy") than with intermediate values of t (high "entropy") there is, on average, a strong intrinsic a-priori bias towards individual functions with low entropy. Furthermore, within a class of functions with fixed t

, we often observe a further intrinsic bias towards functions of lower complexit
y.
Finally, we prove that, regardless of the distribution of inputs, the bias towar
ds low entropy becomes monotonically stronger upon adding ReLU layers, and empir
ically show that increasing the variance of the bias term has a similar effect.
****

DUAL ADVERSARIAL MODEL FOR GENERATING 3D POINT CLOUD
Yuhang Zhang,Zhenwei Miao,Tiebin Mi,Robert Caiming Qiu
Three-dimensional data, such as point clouds, are often composed of three coordi
nates with few featrues. In view of this, it is hard for common neural networks
to learn and represent the characteristics directly. In this paper, we focus on
latent space’s representation of data characteristics, introduce a novel generat
ive framework based on AutoEncoder(AE) and Generative Adversarial Network(GAN) w
ith extra well-designed loss. We embed this framework directly into the raw 3D-G
AN, and experiments demonstrate the potential of the framework in regard of impr
oving the performance on the public dataset compared with other point cloud gene
ration models proposed in recent years. It even achieves state of-the-art perfor
mance. We also perform experiments on MNIST and exhibit an excellent result on 2
D dataset.
****

Wider Networks Learn Better Features
Dar Gilboa,Guy Gur-Ari
Transferability of learned features between tasks can massively reduce the cost
of training a neural network on a novel task. We investigate the effect of netwo
rk width on learned features using activation atlases --- a visualization techni
que that captures features the entire hidden state responds to, as opposed to in
dividual neurons alone. We find that, while individual neurons do not learn inte
rpretable features in wide networks, groups of neurons do. In addition, the hidd
en state of a wide network contains more information about the inputs than that
of a narrow network trained to the same test accuracy. Inspired by this observat
ion, we show that when fine-tuning the last layer of a network on a new task, pe
rformance improves significantly as the width of the network is increased, even
though test accuracy on the original task is independent of width.
****

Conditional Invertible Neural Networks for Guided Image Generation
Lynton Ardizzone,Carsten Lüth,Jakob Kruse,Carsten Rother,Ullrich Köthe
In this work, we address the task of natural image generation guided by a condit
ioning input. We introduce a new architecture called conditional invertible neur
al network (cINN). It combines the purely generative INN model with an unconstra
ined feed-forward network, which efficiently pre-processes the conditioning inpu
t into useful features. All parameters of a cINN are jointly optimized with a s
table, maximum likelihood-based training procedure. Even though INNs and other n
ormalizing flow models have received very little attention in the literature in
contrast to GANs, we find that cINNs can achieve comparable quality, with some r
emarkable properties absent in cGANs, e.g. apparent immunity to mode collapse.
We demonstrate these properties for the tasks of MNIST digit generation and imag
e colorization. Furthermore, we take advantage of our bidirectional cINN archite
cture to explore and manipulate emergent properties of the latent space, such as
 changing the image style in an intuitive way.
****

Cost-Effective Testing of a Deep Learning Model through Input Reduction
Jianyi Zhou,Feng Li,Jinhao Dong,Hongyu Zhang,Dan Hao
With the increasing adoption of Deep Learning (DL) models in various application
s, testing DL models is vitally important. However, testing DL models is costly
and expensive, especially when developers explore alternative designs of DL mode
ls and tune the hyperparameters. To reduce testing cost, we propose to use only
a selected subset of testing data, which is small but representative enough for
quick estimation of the performance of DL models. Our approach, called DeepReduc
e, adopts a two-phase strategy. At first, our approach selects testing data for
the purpose of satisfying testing adequacy. Then, it selects more testing data i

n order to approximate the distribution between the whole testing data and the s
elected data leveraging relative entropy minimization.
Experiments with various DL models and datasets show that our approach can reduc
e the whole testing data to 4.6\% on average, and can reliably estimate the perf
ormance of DL models. Our approach significantly outperforms the random approach
, and is more stable and reliable than the state-of-the-art approach.
****

Hebbian Graph Embeddings
Shalin Shah,Venkataramana Kini
Representation learning has recently been successfully used to create vector rep
resentations of entities in language learning, recommender systems and in simila
rity learning. Graph embeddings exploit the locality structure of a graph and ge
nerate embeddings for nodes which could be words in a language, products on a re
tail website; and the nodes are connected based on a context window. In this pa
per, we consider graph embeddings with an error-free associative learning update
 rule, which models the embedding vector of node as a non-convex Gaussian mixtur
e of the embeddings of the nodes in its immediate vicinity with some constant va
riance that is reduced as iterations progress. It is very easy to parallelize o
ur algorithm without any form of shared memory, which makes it possible to use i
t on very large graphs with a much higher dimensionality of the embeddings. We s
tudy the efficacy of proposed method on several benchmark data sets in Goyal & F
errara(2018b) and favorably compare with state of the art methods. Further, prop
osed method is applied to generate relevant recommendations for a large retailer
.
****

NeuralUCB: Contextual Bandits with Neural Network-Based Exploration
Dongruo Zhou,Lihong Li,Quanquan Gu
We study the stochastic contextual bandit problem, where the reward is generated
 from an unknown bounded function with additive noise. We propose the NeuralUCB
algorithm, which leverages the representation power of deep neural networks and
uses the neural network-based random feature mapping to construct an upper confi
dence bound (UCB) of reward for efficient exploration. We prove that, under mild
 assumptions, NeuralUCB achieves $\tilde O(\sqrt{T})$ regret bound, where T is
 the number of rounds. To the best of our knowledge, our algorithm is the first
neural network-based contextual bandit algorithm with near-optimal regret guaran
tee. Preliminary experiment results on synthetic data corroborate our theory, a
nd shed light on potential applications of our algorithm to real-world problems.
****

Meta-Graph: Few shot Link Prediction via Meta Learning
Avishek Joey Bose,Ankit Jain,Piero Molino,William L. Hamilton
We consider the task of few shot link prediction, where the goal is to predict m
issing edges across multiple graphs using only a small sample of known edges. We
 show that current link prediction methods are generally ill-equipped to handle
this task---as they cannot effectively transfer knowledge between graphs in a mu
lti-graph setting and are unable to effectively learn from very sparse data. To
address this challenge, we introduce a new gradient-based meta learning framewor
k, Meta-Graph, that leverages higher-order gradients along with a learned graph
signature function that conditionally generates a graph neural network initializ
ation. Using a novel set of few shot link prediction benchmarks, we show that Me
ta-Graph enables not only fast adaptation but also better final convergence and
can effectively learn using only a small sample of true edges.
****

Actor-Critic Provably Finds Nash Equilibria of Linear-Quadratic Mean-Field Games
Zuyue Fu,Zhuoran Yang,Yongxin Chen,Zhaoran Wang
We study discrete-time mean-field Markov games with infinite numbers of agents w
here each agent aims to minimize its ergodic cost. We consider the setting where
 the agents have identical linear state transitions and quadratic cost func- tio
ns, while the aggregated effect of the agents is captured by the population mean
 of their states, namely, the mean-field state. For such a game, based on the Na
sh certainty equivalence principle, we provide sufficient conditions for the exi

stence and uniqueness of its Nash equilibrium. Moreover, to find the Nash equili brium, we propose a mean-field actor-critic algorithm with linear function appro xima- tion, which does not require knowing the model of dynamics. Specifically, at each iteration of our algorithm, we use the single-agent actor-critic algorit hm to approximately obtain the optimal policy of the each agent given the curren t mean- field state, and then update the mean-field state. In particular, we pro ve that our algorithm converges to the Nash equilibrium at a linear rate. To the best of our knowledge, this is the first success of applying model-free reinfor cement learn- ing with function approximation to discrete-time mean-field Markov games with provable non-asymptotic global convergence guarantees.
****

An implicit function learning approach for parametric modal regression
Yangchen Pan,Martha White,Amir-massoud Farahmand
For multi-valued functions---such as when the conditional distribution on target s given the inputs is multi-modal---standard regression approaches are not alway s desirable because they provide the conditional mean. Modal regression approach es aim to instead find the conditional mode, but are restricted to nonparametric approaches. Such approaches can be difficult to scale, and make it difficult to benefit from parametric function approximation, like neural networks, which can learn complex relationships between inputs and targets. In this work, we propos e a parametric modal regression algorithm, by using the implicit function theore m to develop an objective for learning a joint parameterized function over input s and targets. We empirically demonstrate on several synthetic problems that our method (i) can learn multi-valued functions and produce the conditional modes, (ii) scales well to high-dimensional inputs and (iii) is even more effective for certain unimodal problems, particularly for high frequency data where the joint function over inputs and targets can better capture the complex relationship be tween them. We conclude by showing that our method provides small improvements o n two regression datasets that have asymmetric distributions over the targets.
****

The asymptotic spectrum of the Hessian of DNN throughout training
Arthur Jacot,Franck Gabriel,Clement Hongler
The dynamics of DNNs during gradient descent is described by the so-called Neura l Tangent Kernel (NTK). In this article, we show that the NTK allows one to gain precise insight into the Hessian of the cost of DNNs: we obtain a full characte rization of the asymptotics of the spectrum of the Hessian, at initialization an d during training.
****

RISE and DISE: Two Frameworks for Learning from Time Series with Missing Data
Alberto Garcia-Duran,Robert West
Time series with missing data constitute an important setting for machine learni ng. The most successful prior approaches for modeling such time series are based on recurrent neural networks that learn to impute unobserved values and then tr eat the imputed values as observed. We start by introducing Recursive Input and State Estimation (RISE), a general framework that encompasses such prior approac hes as specific instances. Since RISE instances tend to suffer from poor long-te rm performance as errors are amplified in feedback loops, we propose Direct Inpu t and State Estimation (DISE), a novel framework in which input and state repres entations are learned from observed data only. The key to DISE is to include tim e information in representation learning, which enables the direct modeling of a rbitrary future time steps by effectively skipping over missing values, rather t han imputing them, thus overcoming the error amplification encountered by RISE m ethods. We benchmark instances of both frameworks on two forecasting tasks, obse rving that DISE achieves state-of-the-art performance on both.
****

Fast Machine Learning with Byzantine Workers and Servers
El-Mahdi El-Mhamdi,Rachid Guerraoui,Arsany Guirguis
Machine Learning (ML) solutions are nowadays distributed and are prone to variou s types of component failures, which can be encompassed in so-called Byzantine b ehavior. This paper introduces LiuBei, a Byzantine-resilient ML algorithm that d

oes not trust any individual component in the network (neither workers nor servers), nor does it induce additional communication rounds (on average), compared to standard non-Byzantine resilient algorithms. LiuBei builds upon gradient aggregation rules (GARs) to tolerate a minority of Byzantine workers. Besides, LiuBei replicates the parameter server on multiple machines instead of trusting it. We introduce a novel filtering mechanism that enables workers to filter out replies from Byzantine server replicas without requiring communication with all servers. Such a filtering mechanism is based on network synchrony, Lipschitz continuity of the loss function, and the GAR used to aggregate workers' gradients. We also introduce a protocol, scatter/gather, to bound drifts between models on correct servers with a small number of communication messages. We theoretically prove that LiuBei achieves Byzantine resilience to both servers and workers and guarantees convergence. We build LiuBei using TensorFlow, and we show that LiuBei tolerates Byzantine behavior with an accuracy loss of around 5% and around 24% convergence overhead compared to vanilla TensorFlow. We moreover show that the throughput gain of LiuBei compared to another state–of–the–art Byzantine–resilient ML algorithm (that assumes network asynchrony) is 70%.

****

How the Softmax Activation Hinders the Detection of Adversarial and Out-of-Distribution Examples in Neural Networks

Jonathan Aigrain,Marcin Detyniecki

Despite having excellent performances for a wide variety of tasks, modern neural networks are unable to provide a prediction with a reliable confidence estimate which would allow to detect misclassifications. This limitation is at the heart of what is known as an adversarial example, where the network provides a wrong prediction associated with a strong confidence to a slightly modified image. Moreover, this overconfidence issue has also been observed for out-of-distribution data. We show through several experiments that the softmax activation, usually placed as the last layer of modern neural networks, is partly responsible for this behaviour. We give qualitative insights about its impact on the MNIST dataset, showing that relevant information present in the logits is lost once the softmax function is applied. The same observation is made through quantitative analysis, as we show that two out-of-distribution and adversarial example detectors obtain competitive results when using logit values as inputs, but provide considerably lower performances if they use softmax probabilities instead: from 98.0% average AUROC to 56.8% in some settings. These results provide evidence that the softmax activation hinders the detection of adversarial and out-of-distribution examples, as it masks a significant part of the relevant information present in the logits.

****

Tree-Structured Attention with Hierarchical Accumulation

Xuan-Phi Nguyen,Shafiq Joty,Steven Hoi,Richard Socher

Incorporating hierarchical structures like constituency trees has been shown to be effective for various natural language processing (NLP) tasks. However, it is evident that state-of-the-art (SOTA) sequence-based models like the Transformer struggle to encode such structures inherently. On the other hand, dedicated models like the Tree-LSTM, while explicitly modeling hierarchical structures, do not perform as efficiently as the Transformer. In this paper, we attempt to bridge this gap with Hierarchical Accumulation to encode parse tree structures into self-attention at constant time complexity. Our approach outperforms SOTA methods in four IWSLT translation tasks and the WMT'14 English-German task. It also yields improvements over Transformer and Tree-LSTM on three text classification tasks. We further demonstrate that using hierarchical priors can compensate for data shortage, and that our model prefers phrase-level attentions over token-level attentions.

****

Deep 3D Pan via local adaptive "t-shaped" convolutions with global and local adaptive dilations

Juan Luis Gonzalez Bello,Munchurl Kim

Recent advances in deep learning have shown promising results in many low-level

vision tasks. However, solving the single-image-based view synthesis is still an open problem. In particular, the generation of new images at parallel camera views given a single input image is of great interest, as it enables 3D visualization of the 2D input scenery. We propose a novel network architecture to perform stereoscopic view synthesis at arbitrary camera positions along the X-axis, or “Deep 3D Pan”, with “t-shaped” adaptive kernels equipped with globally and locally adaptive dilations. Our proposed network architecture, the monster-net, is devised with a novel t-shaped adaptive kernel with globally and locally adaptive dilation, which can efficiently incorporate global camera shift into and handle local 3D geometries of the target image’s pixels for the synthesis of naturally looking 3D panned views when a 2-D input image is given. Extensive experiments were performed on the KITTI, CityScapes, and our VICLAB_STEREO indoors dataset to prove the efficacy of our method. Our monster-net significantly outperforms the state-of-the-art method (SOTA) by a large margin in all metrics of RMSE, PSNR, and SSIM. Our proposed monster-net is capable of reconstructing more reliable image structures in synthesized images with coherent geometry. Moreover, the disparity information that can be extracted from the “t-shaped” kernel is much more reliable than that of the SOTA for the unsupervised monocular depth estimation task, confirming the effectiveness of our method.

****

MANAS: Multi-Agent Neural Architecture Search

Fabio Maria Carlucci,Pedro M Esperança,Marco Singh,Victor Gabillon,Antoine Yang,Hang Xu,Zewei Chen,Jun Wang

The Neural Architecture Search (NAS) problem is typically formulated as a graph search problem where the goal is to learn the optimal operations over edges in order to maximize a graph-level global objective. Due to the large architecture parameter space, efficiency is a key bottleneck preventing NAS from its practical use. In this paper, we address the issue by framing NAS as a multi-agent problem where agents control a subset of the network and coordinate to reach optimal architectures. We provide two distinct lightweight implementations, with reduced memory requirements ($1/8$th of state-of-the-art), and performances above those of much more computationally expensive methods.

Theoretically, we demonstrate vanishing regrets of the form $\mathcal{O}(\sqrt{T})$, with T being the total number of rounds.

Finally, aware that random search is an (often ignored) effective baseline we perform additional experiments on 3 alternative datasets and 2 network configurations, and achieve favorable results in comparison with this baseline and other competing methods.

****

Enhancing Attention with Explicit Phrasal Alignments

Xuan-Phi Nguyen,Shafiq Joty,Thanh-Tung Nguyen

The attention mechanism is an indispensable component of any state-of-the-art neural machine translation system. However, existing attention methods are often token-based and ignore the importance of phrasal alignments, which are the backbone of phrase-based statistical machine translation. We propose a novel phrase-based attention method to model n-grams of tokens as the basic attention entities, and design multi-headed phrasal attentions within the Transformer architecture to perform token-to-token and token-to-phrase mappings. Our approach yields improvements in English-German, English-Russian and English-French translation tasks on the standard WMT'14 test set. Furthermore, our phrasal attention method shows improvements on the one-billion-word language modeling benchmark.

****

LightPAFF: A Two-Stage Distillation Framework for Pre-training and Fine-tuning

Kaitao Song,Hao Sun,Xu Tan,Tao Qin,Jianfeng Lu,Hongzhi Liu,Tie-Yan Liu

While pre-training and fine-tuning, e.g., BERT~\citep{devlin2018bert}, GPT-2~\citep{radford2019language}, have achieved great success in language understanding and generation tasks, the pre-trained models are usually too big for online deployment in terms of both memory cost and inference speed, which hinders them from practical online usage. In this paper, we propose LightPAFF, a Lightweight Pre-

training And Fine-tuning Framework that leverages two-stage knowledge distillation to transfer knowledge from a big teacher model to a lightweight student model in both pre-training and fine-tuning stages. In this way the lightweight model can achieve similar accuracy as the big teacher model, but with much fewer parameters and thus faster online inference speed. LightPAFF can support different pre-training methods (such as BERT, GPT-2 and MASS~\citep{song2019mass}) and be applied to many downstream tasks. Experiments on three language understanding tasks, three language modeling tasks and three sequence to sequence generation tasks demonstrate that while achieving similar accuracy with the big BERT, GPT-2 and MASS models, LightPAFF reduces the model size by nearly 5x and improves online inference speed by 5x-7x.
****
Robust saliency maps with distribution-preserving decoys
Yang Young Lu,Wenbo Guo,Xinyu Xing,William Stafford Noble
Saliency methods help to make deep neural network predictions more interpretable by identifying particular features, such as pixels in an image, that contribute most strongly to the network's prediction. Unfortunately, recent evidence suggests that many saliency methods perform poorly when gradients are saturated or in the presence of strong inter-feature dependence or noise injected by an adversarial attack. In this work, we propose a data-driven technique that uses the distribution-preserving decoys to infer robust saliency scores in conjunction with a pre-trained convolutional neural network classifier and any off-the-shelf saliency method. We formulate the generation of decoys as an optimization problem, potentially applicable to any convolutional network architecture. We also propose a novel decoy-enhanced saliency score, which provably compensates for gradient saturation and considers joint activation patterns of pixels in a single-layer convolutional neural network. Empirical results on the ImageNet data set using three different deep neural network architectures---VGGNet, AlexNet and ResNet---show both qualitatively and quantitatively that decoy-enhanced saliency scores outperform raw scores produced by three existing saliency methods.
****
Role of two learning rates in convergence of model-agnostic meta-learning
Shiro Takagi,Yoshihiro Nagano,Yuki Yoshida,Masato Okada
Model-agnostic meta-learning (MAML) is known as a powerful meta-learning method. However, MAML is notorious for being hard to train because of the existence of two learning rates. Therefore, in this paper, we derive the conditions that inner learning rate α and meta-learning rate β must satisfy for MAML to converge to minima with some simplifications. We find that the upper bound of β depends on $ \alpha$, in contrast to the case of using the normal gradient descent method. Moreover, we show that the threshold of β increases as α approaches its own upper bound. This result is verified by experiments on various few-shot tasks and architectures; specifically, we perform sinusoid regression and classification of Omniglot and MiniImagenet datasets with a multilayer perceptron and a convolutional neural network. Based on this outcome, we present a guideline for determining the learning rates: first, search for the largest possible α; next, tune β based on the chosen value of α.
****
Low-Resource Knowledge-Grounded Dialogue Generation
Xueliang Zhao,Wei Wu,Chongyang Tao,Can Xu,Dongyan Zhao,Rui Yan
Responding with knowledge has been recognized as an important capability for an intelligent conversational agent. Yet knowledge-grounded dialogues, as training data for learning such a response generation model, are difficult to obtain. Motivated by the challenge in practice, we consider knowledge-grounded dialogue generation under a natural assumption that only limited training examples are available. In such a low-resource setting, we devise a disentangled response decoder in order to isolate parameters that depend on knowledge-grounded dialogues from the entire generation model. By this means, the major part of the model can be learned from a large number of ungrounded dialogues and unstructured documents, while the remaining small parameters can be well fitted using the limited training examples. Evaluation results on two benchmarks indicate that with only $1/8$ t

raining data, our model can achieve the state-of-the-art performance and general ize well on out-of-domain knowledge.
****
GResNet: Graph Residual Network for Reviving Deep GNNs from Suspended Animation
Jiawei Zhang,Lin Meng
The existing graph neural networks (GNNs) based on the spectral graph convolutio nal operator have been criticized for its performance degradation, which is espe cially common for the models with deep architectures. In this paper, we further identify the suspended animation problem with the existing GNNs. Such a problem happens when the model depth reaches the suspended animation limit, and the mode l will not respond to the training data any more and become not learnable. Analy sis about the causes of the suspended animation problem with existing GNNs will be provided in this paper, whereas several other peripheral factors that will im pact the problem will be reported as well. To resolve the problem, we introduce the GRESNET (Graph Residual Network) framework in this paper, which creates exte nsively connected highways to involve nodes' raw features or intermediate repres entations throughout the graph for all the model layers. Different from the othe r learning settings, the extensive connections in the graph data will render the existing simple residual learning methods fail to work. We prove the effectiven ess of the introduced new graph residual terms from the norm preservation perspe ctive, which will help avoid dramatic changes to the node's representations betw een sequential layers. Detailed studies about the GRESNET framework for many exi sting GNNs, including GCN, GAT and LOOPYNET, will be reported in the paper with extensive empirical experiments on real-world benchmark datasets.
****
Realism Index: Interpolation in Generative Models With Arbitrary Prior
■ukasz Struski,Jacek Tabor,Igor Podolak,Aleksandra Nowak,Krzysztof Maziarz
In order to perform plausible interpolations in the latent space of a generative model, we need a measure that credibly reflects if a point in an interpolation is close to the data manifold being modelled, i.e. if it is convincing. In this paper, we introduce a realism index of a point, which can be constructed from an arbitrary prior density, or based on FID score approach in case a prior is not available. We propose a numerically efficient algorithm that directly maximises the realism index of an interpolation which, as we theoretically prove, leads to a search of a geodesic with respect to the corresponding Riemann structure. We show that we obtain better interpolations then the classical linear ones, in par ticular when either the prior density is not convex shaped, or when the soap bub ble effect appears.
****
Deep RL for Blood Glucose Control: Lessons, Challenges, and Opportunities
Ian Fox,Joyce Lee,Rodica Busui,Jenna Wiens
Individuals with type 1 diabetes (T1D) lack the ability to produce the insulin t heir bodies need. As a result, they must continually make decisions about how mu ch insulin to self-administer in order to adequately control their blood glucose levels. Longitudinal data streams captured from wearables, like continuous gluc ose monitors, can help these individuals manage their health, but currently the majority of the decision burden remains on the user. To relieve this burden, res earchers are working on closed-loop solutions that combine a continuous glucose monitor and an insulin pump with a control algorithm in an `artificial pancreas. ' Such systems aim to estimate and deliver the appropriate amount of insulin. He re, we develop reinforcement learning (RL) techniques for automated blood glucos e control. Through a series of experiments, we compare the performance of differ ent deep RL approaches to non-RL approaches. We highlight the flexibility of RL approaches, demonstrating how they can adapt to new individuals with little addi tional data. On over 21k hours of simulated data across 30 patients, RL approach es outperform baseline control algorithms (increasing time spent in normal gluco se range from 71% to 75%) without requiring meal announcements. Moreover, these approaches are adept at leveraging latent behavioral patterns (increasing time i n range from 58% to 70%). This work demonstrates the potential of deep RL for co ntrolling complex physiological systems with minimal expert knowledge.

****
A Target-Agnostic Attack on Deep Models: Exploiting Security Vulnerabilities of Transfer Learning
Shahbaz Rezaei,Xin Liu
Due to insufficient training data and the high computational cost to train a deep neural network from scratch, transfer learning has been extensively used in many deep-neural-network-based applications. A commonly used transfer learning approach involves taking a part of a pre-trained model, adding a few layers at the end, and re-training the new layers with a small dataset. This approach, while efficient and widely used, imposes a security vulnerability because the pre-trained model used in transfer learning is usually publicly available, including to potential attackers. In this paper, we show that without any additional knowledge other than the pre-trained model, an attacker can launch an effective and efficient brute force attack that can craft instances of input to trigger each target class with high confidence. We assume that the attacker has no access to any target-specific information, including samples from target classes, re-trained model, and probabilities assigned by Softmax to each class, and thus making the attack target-agnostic. These assumptions render all previous attack models inapplicable, to the best of our knowledge. To evaluate the proposed attack, we perform a set of experiments on face recognition and speech recognition tasks and show the effectiveness of the attack. Our work reveals a fundamental security weakness of the Softmax layer when used in transfer learning settings.
****
Training Provably Robust Models by Polyhedral Envelope Regularization
Chen Liu,Mathieu Salzmann,Sabine Süsstrunk
Training certifiable neural networks enables one to obtain models with robustness guarantees against adversarial attacks. In this work, we use a linear approximation to bound model’s output given an input adversarial budget. This allows us to bound the adversary-free region in the data neighborhood by a polyhedral envelope and yields finer-grained certified robustness than existing methods. We further exploit this certifier to introduce a framework called polyhedral envelope regular- ization (PER), which encourages larger polyhedral envelopes and thus improves the provable robustness of the models. We demonstrate the flexibility and effectiveness of our framework on standard benchmarks; it applies to networks with general activation functions and obtains comparable or better robustness guarantees than state-of-the-art methods, with very little cost in clean accuracy, i.e., without over-regularizing the model.
****
FleXOR: Trainable Fractional Quantization
Dongsoo Lee,Se Jung Kwon,Byeongwook Kim,Yongkweon Jeon,Baeseong Park,Jeongin Yun,Gu-Yeon Wei
Parameter quantization is a popular model compression technique due to its regular form and high compression ratio. In particular, quantization based on binary codes is gaining attention because each quantized bit can be directly utilized for computations without dequantization using look-up tables. Previous attempts, however, only allow for integer numbers of quantization bits, which ends up restricting the search space for compression ratio and accuracy. Moreover, quantization bits are usually obtained by minimizing quantization loss in a local manner that does not directly correspond to minimizing the loss function. In this paper, we propose an encryption algorithm/architecture to compress quantized weights in order to achieve fractional numbers of bits per weight and new compression configurations further optimize accuracy/compression trade-offs. Decryption is implemented using XOR gates added into the neural network model and described as $\tanh(x)$, which enable gradient calculations superior to the straight-through gradient method. We perform experiments using MNIST, CIFAR-10, and ImageNet to show that inserting XOR gates learns quantization/encrypted bit decisions through training and obtains high accuracy even for fractional sub 1-bit weights.
****
Multi-Task Learning via Scale Aware Feature Pyramid Networks and Effective Joint Head

Feng Ni
As a concise and classic framework for object detection and instance segmentatio n, Mask R-CNN achieves promising performance in both two tasks. However, conside ring stronger feature representation for Mask R-CNN fashion framework, there is room for improvement from two aspects. On the one hand, performing multi-task pr ediction needs more credible feature extraction and multi-scale features integra tion to handle objects with varied scales. In this paper, we address this proble m by using a novel neck module called SA-FPN (Scale Aware Feature Pyramid Networ ks). With the enhanced feature representations, our model can accurately detect and segment the objects of multiple scales. On the other hand, in Mask R-CNN fra mework, isolation between parallel detection branch and instance segmentation br anch exists, causing the gap between training and testing processes. To narrow t his gap, we propose a unified head module named EJ-Head (Effective Joint Head) t o combine two branches into one head, not only realizing the interaction between two tasks, but also enhancing the effectiveness of multi-task learning. Compreh ensive experiments show that our proposed methods bring noticeable gains for obj ect detection and instance segmentation. In particular, our model outperforms th e original Mask R-CNN by 1~2 percent AP in both object detection and instance se gmentation task on MS-COCO benchmark. Code will be available soon.
****

AdaX: Adaptive Gradient Descent with Exponential Long Term Memory
Wenjie Li,Zhaoyang Zhang,Xinjiang Wang,Ping Luo
Adaptive optimization algorithms such as RMSProp and Adam have fast convergence and smooth learning process. Despite their successes, they are proven to have no n-convergence issue even in convex optimization problems as well as weak perform ance compared with the first order gradient methods such as stochastic gradient descent (SGD). Several other algorithms, for example AMSGrad and AdaShift, have been proposed to alleviate these issues but only minor effect has been observed. This paper further analyzes the performance of such algorithms in a non-convex setting by extending their non-convergence issue into a simple non-convex case a nd show that Adam's design of update steps would possibly lead the algorithm to local minimums. To address the above problems, we propose a novel adaptive gradi ent descent algorithm, named AdaX, which accumulates the long-term past gradient information exponentially. We prove the convergence of AdaX in both convex and non-convex settings. Extensive experiments show that AdaX outperforms Adam in va rious tasks of computer vision and natural language processing and can catch up with SGD.

****

On Computation and Generalization of Generative Adversarial Imitation Learning
Minshuo Chen,Yizhou Wang,Tianyi Liu,Zhuoran Yang,Xingguo Li,Zhaoran Wang,Tuo Zha o
Generative Adversarial Imitation Learning (GAIL) is a powerful and practical app roach for learning sequential decision-making policies. Different from Reinforce ment Learning (RL), GAIL takes advantage of demonstration data by experts (e.g., human), and learns both the policy and reward function of the unknown environme nt. Despite the significant empirical progresses, the theory behind GAIL is stil l largely unknown. The major difficulty comes from the underlying temporal depen dency of the demonstration data and the minimax computational formulation of GAI L without convex-concave structure. To bridge such a gap between theory and prac tice, this paper investigates the theoretical properties of GAIL. Specifically, we show: (1) For GAIL with general reward parameterization, the generalization c an be guaranteed as long as the class of the reward functions is properly contro lled; (2) For GAIL, where the reward is parameterized as a reproducing kernel fu nction, GAIL can be efficiently solved by stochastic first order optimization al gorithms, which attain sublinear convergence to a stationary solution. To the be st of our knowledge, these are the first results on statistical and computationa l guarantees of imitation learning with reward/policy function ap- proximation. Numerical experiments are provided to support our analysis.

****
Disentangling Improves VAEs' Robustness to Adversarial Attacks
Matthew Willetts,Alexander Camuto,Stephen Roberts,Chris Holmes
This paper is concerned with the robustness of VAEs to adversarial attacks. We highlight that conventional VAEs are brittle under attack but that methods recently introduced for disentanglement such as β-TCVAE (Chen et al., 2018) improve robustness, as demonstrated through a variety of previously proposed adversarial attacks (Tabacof et al. (2016); Gondim-Ribeiro et al. (2018); Kos et al.(2018)). This motivated us to develop Seatbelt-VAE, a new hierarchical disentangled VAE that is designed to be significantly more robust to adversarial attacks than existing approaches, while retaining high quality reconstructions.
****
Sparsity Meets Robustness: Channel Pruning for the Feynman-Kac Formalism Principled Robust Deep Neural Nets
Thu Dinh*,Bao Wang*,Andrea L. Bertozzi,Stanley J. Osher,Jack Xin
Deep neural nets (DNNs) compression is crucial for adaptation to mobile devices. Though many successful algorithms exist to compress naturally trained DNNs, developing efficient and stable compression algorithms for robustly trained DNNs remains widely open. In this paper, we focus on a co-design of efficient DNN compression algorithms and sparse neural architectures for robust and accurate deep learning. Such a co-design enables us to advance the goal of accommodating both sparsity and robustness. With this objective in mind, we leverage the relaxed augmented Lagrangian based algorithms to prune the weights of adversarially trained DNNs, at both structured and unstructured levels. Using a Feynman-Kac formalism principled robust and sparse DNNs, we can at least double the channel sparsity of the adversarially trained ResNet20 for CIFAR10 classification, meanwhile, improve the natural accuracy by 8.69\% and the robust accuracy under the benchmark 20 iterations of IFGSM attack by 5.42\%.
****
FEW-SHOT LEARNING ON GRAPHS VIA SUPER-CLASSES BASED ON GRAPH SPECTRAL MEASURES
Jatin Chauhan,Deepak Nathani,Manohar Kaul
We propose to study the problem of few-shot graph classification in graph neural networks (GNNs) to recognize unseen classes, given limited labeled graph examples. Despite several interesting GNN variants being proposed recently for node and graph classification tasks, when faced with scarce labeled examples in the few-shot setting, these GNNs exhibit significant loss in classification performance. Here, we present an approach where a probability measure is assigned to each graph based on the spectrum of the graph’s normalized Laplacian. This enables us to accordingly cluster the graph base-labels associated with each graph into super-classes, where the L^p Wasserstein distance serves as our underlying distance metric. Subsequently, a super-graph constructed based on the super-classes is then fed to our proposed GNN framework which exploits the latent inter-class relationships made explicit by the super-graph to achieve better class label separation among the graphs. We conduct exhaustive empirical evaluations of our proposed method and show that it outperforms both the adaptation of state-of-the-art graph classification methods to few-shot scenario and our naive baseline GNNs. Additionally, we also extend and study the behavior of our method to semi-supervised and active learning scenarios.
****
Influence-Based Multi-Agent Exploration
Tonghan Wang*,Jianhao Wang*,Yi Wu,Chongjie Zhang
Intrinsically motivated reinforcement learning aims to address the exploration challenge for sparse-reward tasks. However, the study of exploration methods in transition-dependent multi-agent settings is largely absent from the literature. We aim to take a step towards solving this problem. We present two exploration methods: exploration via information-theoretic influence (EITI) and exploration via decision-theoretic influence (EDTI), by exploiting the role of interaction in coordinated behaviors of agents. EITI uses mutual information to capture the interdependence between the transition dynamics of agents. EDTI uses a novel intrinsic reward, called Value of Interaction (VoI), to characterize and quantify the

influence of one agent's behavior on expected returns of other agents. By optimizing EITI or EDTI objective as a regularizer, agents are encouraged to coordinate their exploration and learn policies to optimize the team performance. We show how to optimize these regularizers so that they can be easily integrated with policy gradient reinforcement learning. The resulting update rule draws a connection between coordinated exploration and intrinsic reward distribution. Finally, we empirically demonstrate the significant strength of our methods in a variety of multi-agent scenarios.
****
Demonstration Actor Critic
Guoqing Liu,Li Zhao,Pushi Zhang,Jiang Bian,Tao Qin,Nenghai Yu,Tie-Yan Liu
We study the problem of \textit{Reinforcement learning from demonstrations (RLfD)}, where the learner is provided with both some expert demonstrations and reinforcement signals from the environment. One approach leverages demonstration data in a supervised manner, which is simple and direct, but can only provide supervision signal over those states seen in the demonstrations. Another approach uses demonstration data for reward shaping. By contrast, the latter approach can provide guidance on how to take actions, even for those states are not seen in the demonstrations. But existing algorithms in the latter one adopt shaping reward which is not directly dependent on current policy, limiting the algorithms to treat demonstrated states the same as other states, failing to directly exploit supervision signal in demonstration data. In this paper, we propose a novel objective function with policy-dependent shaping reward, so as to get the best of both worlds. We present a convergence proof for policy iteration of the proposed objective, under the tabular setting. Then we develop a new practical algorithm, termed as Demonstration Actor Critic (DAC). Experiments on a range of popular benchmark sparse-reward tasks shows that our DAC method obtains a significant performance gain over five strong and off-the-shelf baselines.
****
Deep Coordination Graphs
Wendelin Boehmer,Vitaly Kurin,Shimon Whiteson
This paper introduces the deep coordination graph (DCG) for collaborative multi-agent reinforcement learning. DCG strikes a flexible trade-off between representational capacity and generalization by factorizing the joint value function of all agents according to a coordination graph into payoffs between pairs of agents. The value can be maximized by local message passing along the graph, which allows training of the value function end-to-end with Q-learning. Payoff functions are approximated with deep neural networks and parameter sharing improves generalization over the state-action space. We show that DCG can solve challenging predator-prey tasks that are vulnerable to the relative overgeneralization pathology and in which all other known value factorization approaches fail.
****
Cross-Dimensional Self-Attention for Multivariate, Geo-tagged Time Series Imputation
Jiawei Ma*,Zheng Shou*,Alireza Zareian,Hassan Mansour,Anthony Vetro,Shih-Fu Chang
Many real-world applications involve multivariate, geo-tagged time series data: at each location, multiple sensors record corresponding measurements. For example, air quality monitoring system records PM2.5, CO, etc. The resulting time-series data often has missing values due to device outages or communication errors. In order to impute the missing values, state-of-the-art methods are built on Recurrent Neural Networks (RNN), which process each time stamp sequentially, prohibiting the direct modeling of the relationship between distant time stamps. Recently, the self-attention mechanism has been proposed for sequence modeling tasks such as machine translation, significantly outperforming RNN because the relationship between each two time stamps can be modeled explicitly. In this paper, we are the first to adapt the self-attention mechanism for multivariate, geo-tagged time series data. In order to jointly capture the self-attention across different dimensions (i.e. time, location and sensor measurements) while keep the size of attention maps reasonable, we propose a novel approach called Cross-Dimension

al Self-Attention (CDSA) to process each dimension sequentially, yet in an order -independent manner. On three real-world datasets, including one our newly colle cted NYC-traffic dataset, extensive experiments demonstrate the superiority of o ur approach compared to state-of-the-art methods for both imputation and forecas ting tasks.

****

How Well Do WGANs Estimate the Wasserstein Metric?
Anton Mallasto,Guido Montúfar,Augusto Gerolin
Generative modelling is often cast as minimizing a similarity measure between a data distribution and a model distribution. Recently, a popular choice for the s imilarity measure has been the Wasserstein metric, which can be expressed in the Kantorovich duality formulation as the optimum difference of the expected value s of a potential function under the real data distribution and the model hypothe sis. In practice, the potential is approximated with a neural network and is cal led the discriminator. Duality constraints on the function class of the discrimi nator are enforced approximately, and the expectations are estimated from sample s. This gives at least three sources of errors: the approximated discriminator a nd constraints, the estimation of the expectation value, and the optimization re quired to find the optimal potential. In this work, we study how well the method s, that are used in generative adversarial networks to approximate the Wasserste in metric, perform. We consider, in particular, the c-transform formulation, w hich eliminates the need to enforce the constraints explicitly. We demonstrate t hat the c-transform allows for a more accurate estimation of the true Wasserst ein metric from samples, but surprisingly, does not
****

Revisiting the Generalization of Adaptive Gradient Methods
Naman Agarwal,Rohan Anil,Elad Hazan,Tomer Koren,Cyril Zhang
A commonplace belief in the machine learning community is that using adaptive gr adient methods hurts generalization. We re-examine this belief both theoreticall y and experimentally, in light of insights and trends from recent years.
We revisit some previous oft-cited experiments and theoretical accounts in more depth, and provide a new set of experiments in larger-scale, state-of-the-art se ttings. We conclude that with proper tuning, the improved training performance o f adaptive optimizers does not in general carry an overfitting penalty, especial ly in contemporary deep learning. Finally, we synthesize a ``user's guide'' to a daptive optimizers, including some proposed modifications to AdaGrad to mitigate some of its empirical shortcomings.
****

Multiplicative Interactions and Where to Find Them
Siddhant M. Jayakumar,Wojciech M. Czarnecki,Jacob Menick,Jonathan Schwarz,Jack R ae,Simon Osindero,Yee Whye Teh,Tim Harley,Razvan Pascanu
We explore the role of multiplicative interaction as a unifying framework to des cribe a range of classical and modern neural network architectural motifs, such as gating, attention layers, hypernetworks, and dynamic convolutions amongst oth ers.
Multiplicative interaction layers as primitive operations have a long-establishe d presence in the literature, though this often not emphasized and thus under-ap preciated. We begin by showing that such layers strictly enrich the representabl e function classes of neural networks. We conjecture that multiplicative interac tions offer a particularly powerful inductive bias when fusing multiple streams of information or when conditional computation is required. We therefore argue t hat they should be considered in many situation where multiple compute or inform ation paths need to be combined, in place of the simple and oft-used concatenati on operation. Finally, we back up our claims and demonstrate the potential of mu ltiplicative interactions by applying them in large-scale complex RL and sequenc e modelling tasks, where their use allows us to deliver state-of-the-art results , and thereby provides new evidence in support of multiplicative interactions pl aying a more prominent role when designing new neural network architectures.
****

SELF-KNOWLEDGE DISTILLATION ADVERSARIAL ATTACK
Ma Xiaoxiong[1],Wang Renzhi[1],Tian Cong,Dong Zeqian,Duan Zhenhua
Neural networks show great vulnerability under the threat of adversarial example s.

By adding small perturbation to a clean image, neural networks with high clas sification accuracy can be completely fooled.

One intriguing property of the adversarial examples is transferability. This property allows adversarial examples to transfer to networks of unknown structu re, which is harmful even to the physical world.

The current way of generating adversarial examples is mainly divided into opt imization based and gradient based methods.

Liu et al. (2017) conjecture that gradient based methods can hardly produce t ransferable targeted adversarial examples in black-box-attack.

However, in this paper, we use a simple technique to improve the transferabil ity and success rate of targeted attacks with gradient based methods.

We prove that gradient based methods can also generate transferable adversari al examples in targeted attacks.

Specifically, we use knowledge distillation for gradient based methods, and s how that the transferability can be improved by effectively utilizing different classes of information.

Unlike the usual applications of knowledge distillation, we did not train a s tudent network to generate adversarial examples.

We take advantage of the fact that knowledge distillation can soften the targ et and obtain higher information, and combine the soft target and hard target of the same network as the loss function.

Our method is generally applicable to most gradient based attack methods.
****
DIVA: Domain Invariant Variational Autoencoder
Maximilian Ilse,Jakub M. Tomczak,Christos Louizos,Max Welling
We consider the problem of domain generalization, namely, how to learn represent ations given data from a set of domains that generalize to data from a previousl y unseen domain. We propose the Domain Invariant Variational Autoencoder (DIVA), a generative model that tackles this problem by learning three independent late nt subspaces, one for the domain, one for the class, and one for any residual va riations. We highlight that due to the generative nature of our model we can als o incorporate unlabeled data from known or previously unseen domains. To the bes t of our knowledge this has not been done before in a domain generalization sett ing. This property is highly desirable in fields like medical imaging where labe led data is scarce. We experimentally evaluate our model on the rotated MNIST be nchmark and a malaria cell images dataset where we show that (i) the learned sub spaces are indeed complementary to each other, (ii) we improve upon recent works on this task and (iii) incorporating unlabelled data can boost the performance even further.
****
Continual Learning with Bayesian Neural Networks for Non-Stationary Data
Richard Kurle,Botond Cseke,Alexej Klushyn,Patrick van der Smagt,Stephan Günneman n
This work addresses continual learning for non-stationary data, using Bayesian n eural networks and memory-based online variational Bayes. We represent the poste rior approximation of the network weights by a diagonal Gaussian distribution an d a complementary memory of raw data. This raw data corresponds to likelihood te rms that cannot be well approximated by the Gaussian. We introduce a novel metho d for sequentially updating both components of the posterior approximation. Furt hermore, we propose Bayesian forgetting and a Gaussian diffusion process for ada pting to non-stationary data. The experimental results show that our update meth od improves on existing approaches for streaming data. Additionally, the adaptat ion methods lead to better predictive performance for non-stationary data.
****
RPGAN: random paths as a latent space for GAN interpretability
Andrey Voynov,Artem Babenko

In this paper, we introduce Random Path Generative Adversarial Network (RPGAN) --- an alternative scheme of GANs that can serve as a tool for generative model analysis. While the latent space of a typical GAN consists of input vectors, randomly sampled from the standard Gaussian distribution, the latent space of RPGAN consists of random paths in a generator network. As we show, this design allows to associate different layers of the generator with different regions of the latent space, providing their natural interpretability. With experiments on standard benchmarks, we demonstrate that RPGAN reveals several interesting insights about roles that different layers play in the image generation process. Aside from interpretability, the RPGAN model also provides competitive generation quality and allows efficient incremental learning on new data.
****
SAdam: A Variant of Adam for Strongly Convex Functions
Guanghui Wang,Shiyin Lu,Quan Cheng,Wei-wei Tu,Lijun Zhang
The Adam algorithm has become extremely popular for large-scale machine learning. Under convexity condition, it has been proved to enjoy a data-dependent $O(\sqrt{T})$ regret bound where T is the time horizon. However, whether strong convexity can be utilized to further improve the performance remains an open problem. In this paper, we give an affirmative answer by developing a variant of Adam (referred to as SAdam) which achieves a data-dependent $O(\log T)$ regret bound for strongly convex functions. The essential idea is to maintain a faster decaying yet under controlled step size for exploiting strong convexity. In addition, under a special configuration of hyperparameters, our SAdam reduces to SC-RMSprop, a recently proposed variant of RMSprop for strongly convex functions, for which we provide the first data-dependent logarithmic regret bound. Empirical results on optimizing strongly convex functions and training deep networks demonstrate the effectiveness of our method.
****
Improving the Generalization of Visual Navigation Policies using Invariance Regularization
Michel Aractingi,Christopher Dance,Julien Perez,Tomi Silander
Training agents to operate in one environment often yields overfitted models that are unable to generalize to the changes in that environment. However, due to the numerous variations that can occur in the real-world, the agent is often required to be robust in order to be useful. This has not been the case for agents trained with reinforcement learning (RL) algorithms. In this paper, we investigate the overfitting of RL agents to the training environments in visual navigation tasks. Our experiments show that deep RL agents can overfit even when trained on multiple environments simultaneously.
We propose a regularization method which combines RL with supervised learning methods by adding a term to the RL objective that would encourage the invariance of a policy to variations in the observations that ought not to affect the action taken. The results of this method, called invariance regularization, show an improvement in the generalization of policies to environments not seen during training.

****
Generalization bounds for deep convolutional neural networks
Philip M. Long,Hanie Sedghi
We prove bounds on the generalization error of convolutional networks.
The bounds are in terms of the training loss, the number of
parameters, the Lipschitz constant of the loss and the distance from
the weights to the initial weights. They are independent of the
number of pixels in the input, and the height and width of hidden
feature maps.
We present experiments using CIFAR-10 with varying
hyperparameters of a deep convolutional network, comparing our bounds
with practical generalization gaps.
****
Scaling Laws for the Principled Design, Initialization, and Preconditioning of R

eLU Networks
Aaron Defazio,Leon Bottou
Abstract In this work, we describe a set of rules for the design and initialization of well-conditioned neural networks, guided by the goal of naturally balancing the diagonal blocks of the Hessian at the start of training. We show how our measure of conditioning of a block relates to another natural measure of conditioning, the ratio of weight gradients to the weights. We prove that for a ReLU-based deep multilayer perceptron, a simple initialization scheme using the geometric mean of the fan-in and fan-out satisfies our scaling rule. For more sophisticated architectures, we show how our scaling principle can be used to guide design choices to produce well-conditioned neural networks, reducing guess-work.
****
A Fair Comparison of Graph Neural Networks for Graph Classification
Federico Errica,Marco Podda,Davide Bacciu,Alessio Micheli
Experimental reproducibility and replicability are critical topics in machine learning. Authors have often raised concerns about their lack in scientific publications to improve the quality of the field. Recently, the graph representation learning field has attracted the attention of a wide research community, which resulted in a large stream of works.
As such, several Graph Neural Network models have been developed to effectively tackle graph classification. However, experimental procedures often lack rigorousness and are hardly reproducible. Motivated by this, we provide an overview of common practices that should be avoided to fairly compare with the state of the art. To counter this troubling trend, we ran more than 47000 experiments in a controlled and uniform framework to re-evaluate five popular models across nine common benchmarks. Moreover, by comparing GNNs with structure-agnostic baselines we provide convincing evidence that, on some datasets, structural information has not been exploited yet. We believe that this work can contribute to the development of the graph learning field, by providing a much needed grounding for rigorous evaluations of graph classification models.
****
Finding and Visualizing Weaknesses of Deep Reinforcement Learning Agents
Christian Rupprecht,Cyril Ibrahim,Christopher J. Pal
As deep reinforcement learning driven by visual perception becomes more widely used there is a growing need to better understand and probe the learned agents. Understanding the decision making process and its relationship to visual inputs can be very valuable to identify problems in learned behavior. However, this topic has been relatively under-explored in the research community. In this work we present a method for synthesizing visual inputs of interest for a trained agent. Such inputs or states could be situations in which specific actions are necessary. Further, critical states in which a very high or a very low reward can be achieved are often interesting to understand the situational awareness of the system as they can correspond to risky states. To this end, we learn a generative model over the state space of the environment and use its latent space to optimize a target function for the state of interest. In our experiments we show that this method can generate insights for a variety of environments and reinforcement learning methods. We explore results in the standard Atari benchmark games as well as in an autonomous driving simulator. Based on the efficiency with which we have been able to identify behavioural weaknesses with this technique, we believe this general approach could serve as an important tool for AI safety applications.
****
Computation Reallocation for Object Detection
Feng Liang,Chen Lin,Ronghao Guo,Ming Sun,Wei Wu,Junjie Yan,Wanli Ouyang
The allocation of computation resources in the backbone is a crucial issue in object detection. However, classification allocation pattern is usually adopted directly to object detector, which is proved to be sub-optimal. In order to reallocate the engaged computation resources in a more efficient way, we present CR-NAS (Computation Reallocation Neural Architecture Search) that can learn computation reallocation strategies across different feature resolution and spatial posit

ion diectly on the target detection dataset. A two-level reallocation space is p
roposed for both stage and spatial reallocation. A novel hierarchical search pro
cedure is adopted to cope with the complex search space. We apply CR-NAS to mult
iple backbones and achieve consistent improvements. Our CR-ResNet50 and CR-Mobil
eNetV2 outperforms the baseline by 1.9% and 1.7% COCO AP respectively without an
y additional computation budget. The models discovered by CR-NAS can be equiped
to other powerful detection neck/head and be easily transferred to other dataset
, e.g. PASCAL VOC, and other vision tasks, e.g. instance segmentation. Our CR-NA
S can be used as a plugin to improve the performance of various networks, which
is demanding.
****
MULTI-LABEL METRIC LEARNING WITH BIDIRECTIONAL REPRESENTATION DEEP NEURAL NETWOR
KS
Tao Zheng,Ivor Tsang,Xin Yao
Multi-Label Learning task simultaneously predicting multiple labels has attracte
d researchers' interest for its wide application.
Metric Learning crucially determines the performance of the k nearest neighbor a
lgorithms, the most popular framework handling the multi-label problem.
However, the existing advanced multiple-label metric learning suffers the inferi
or capacity and application restriction.
We propose an extendable and end-to-end deep representation approach for metric
learning on multi-label data set that is based on neural networks able to operat
e on feature data or directly on raw image data.
We motivate the choice of our network architecture via a Bidirectional Represent
ation learning where the label dependency is also integrated and deep convolutio
nal networks that handle image data.
In multi-label metric learning, instances with the more different labels will be
 dragged the more far away, but ones with identical labels will concentrate toge
ther.
Our model scales linearly in the number of instances and trains deep neural netw
orks that encode both input data and output labels, then, obtains a metric space
 for testing data.
In a number of experiments on multi-labels tasks, we demonstrate that our approa
ch is better than related methods based on the systematic metric and its extenda
bility.

****
Sparse Networks from Scratch: Faster Training without Losing Performance
Tim Dettmers,Luke Zettlemoyer
We demonstrate the possibility of what we call sparse learning: accelerated trai
ning of deep neural networks that maintain sparse weights throughout training wh
ile achieving dense performance levels. We accomplish this by developing sparse
momentum, an algorithm which uses exponentially smoothed gradients (momentum) to
 identify layers and weights which reduce the error efficiently. Sparse momentum
 redistributes pruned weights across layers according to the mean momentum magni
tude of each layer. Within a layer, sparse momentum grows weights according to t
he momentum magnitude of zero-valued weights. We demonstrate state-of-the-art sp
arse performance on MNIST, CIFAR-10, and ImageNet, decreasing the mean error by
a relative 8%, 15%, and 6% compared to other sparse algorithms. Furthermore, we
show that sparse momentum reliably reproduces dense performance levels while pro
viding up to 5.61x faster training. In our analysis, ablations show that the ben
efits of momentum redistribution and growth increase with the depth and size of
the network.
****
Modeling Winner-Take-All Competition in Sparse Binary Projections
Wenye Li
Inspired by the advances in biological science, the study of sparse binary proje
ction models has attracted considerable recent research attention. The models pr
oject dense input samples into a higher-dimensional space and output sparse bina
ry data representations after Winner-Take-All competition, subject to the constr

aint that the projection matrix is also sparse and binary. Following the work al
ong this line, we developed a supervised-WTA model when training samples with bo
th input and output representations are available, from which the optimal projec
tion matrix can be obtained with a simple, efficient yet effective algorithm. We
 further extended the model and the algorithm to an unsupervised setting where o
nly the input representation of the samples is available. In a series of empiric
al evaluation on similarity search tasks, the proposed models reported significa
ntly improved results over the state-of-the-art methods in both search accuracy
and running time. The successful results give us strong confidence that the work
 provides a highly practical tool to real world applications.

****
Laplacian Denoising Autoencoder
Jianbo Jiao,Linchao Bao,Yunchao Wei,Shengfeng He,Honghui Shi,Rynson Lau,Thomas H
uang
While deep neural networks have been shown to perform remarkably well in many ma
chine learning tasks, labeling a large amount of supervised data is usually very
 costly to scale. Therefore, learning robust representations with unlabeled data
 is critical in relieving human effort and vital for many downstream application
s. Recent advances in unsupervised and self-supervised learning approaches for v
isual data benefit greatly from domain knowledge. Here we are interested in a mo
re generic unsupervised learning framework that can be easily generalized to oth
er domains. In this paper, we propose to learn data representations with a novel
 type of denoising autoencoder, where the input noisy data is generated by corru
pting the clean data in gradient domain. This can be naturally generalized to sp
an multiple scales with a Laplacian pyramid representation of the input data. In
 this way, the agent has to learn more robust representations that can exploit t
he underlying data structures across multiple scales. Experiments on several vis
ual benchmarks demonstrate that better representations can be learned with the p
roposed approach, compared to its counterpart with single-scale corruption. Besi
des, we also demonstrate that the learned representations perform well when tran
sferring to other vision tasks.
****
Training Data Distribution Search with Ensemble Active Learning
Kashyap Chitta,Jose M. Alvarez,Elmar Haussmann,Clement Farabet
Deep Neural Networks (DNNs) often rely on very large datasets for training. Give
n the large size of such datasets, it is conceivable that they contain certain s
amples that either do not contribute or negatively impact the DNN's optimization
. Modifying the training distribution in a way that excludes such samples could
provide an effective solution to both improve performance and reduce training ti
me. In this paper, we propose to scale up ensemble Active Learning methods to pe
rform acquisition at a large scale (10k to 500k samples at a time). We do this w
ith ensembles of hundreds of models, obtained at a minimal computational cost by
 reusing intermediate training checkpoints. This allows us to automatically and
efficiently perform a training data distribution search for large labeled datase
ts. We observe that our approach obtains favorable subsets of training data, whi
ch can be used to train more accurate DNNs than training with the entire dataset
. We perform an extensive experimental study of this phenomenon on three image c
lassification benchmarks (CIFAR-10, CIFAR-100 and ImageNet), analyzing the impac
t of initialization schemes, acquisition functions and ensemble configurations.
We demonstrate that data subsets identified with a lightweight ResNet-18 ensembl
e remain effective when used to train deep models like ResNet-101 and DenseNet-1
21. Our results provide strong empirical evidence that optimizing the training d
ata distribution can provide significant benefits on large scale vision tasks.
****
Meta-Learning without Memorization
Mingzhang Yin,George Tucker,Mingyuan Zhou,Sergey Levine,Chelsea Finn
The ability to learn new concepts with small amounts of data is a critical aspec
t of intelligence that has proven challenging for deep learning methods. Meta-le
arning has emerged as a promising technique for leveraging data from previous ta

sks to enable efficient learning of new tasks. However, most meta-learning algor ithms implicitly require that the meta-training tasks be mutually-exclusive, suc h that no single model can solve all of the tasks at once. For example, when cre ating tasks for few-shot image classification, prior work uses a per-task random assignment of image classes to N-way classification labels. If this is not done , the meta-learner can ignore the task training data and learn a single model th at performs all of the meta-training tasks zero-shot, but does not adapt effecti vely to new image classes. This requirement means that the user must take great care in designing the tasks, for example by shuffling labels or removing task i dentifying information from the inputs. In some domains, this makes meta-learnin g entirely inapplicable. In this paper, we address this challenge by designing a meta-regularization objective using information theory that places precedence o n data-driven adaptation. This causes the meta-learner to decide what must be le arned from the task training data and what should be inferred from the task test ing input. By doing so, our algorithm can successfully use data from non-mutuall y-exclusive tasks to efficiently adapt to novel tasks. We demonstrate its applic ability to both contextual and gradient-based meta-learning algorithms, and appl y it in practical settings where applying standard meta-learning has been diffic ult. Our approach substantially outperforms standard meta-learning algorithms in these settings.

****

From Variational to Deterministic Autoencoders

Partha Ghosh,Mehdi S. M. Sajjadi,Antonio Vergari,Michael Black,Bernhard Scholkop f

Variational Autoencoders (VAEs) provide a theoretically-backed and popular fram ework for deep generative models. However, learning a VAE from data poses still unanswered theoretical questions and considerable practical challenges. In this work, we propose an alternative framework for generative modeling that is simple r, easier to train, and deterministic, yet has many of the advantages of the VAE . We observe that sampling a stochastic encoder in a Gaussian VAE can be interpr eted as simply injecting noise into the input of a deterministic decoder. We inv estigate how substituting this kind of stochasticity, with other explicit and im plicit regularization schemes, can lead to an equally smooth and meaningful late nt space without having to force it to conform to an arbitrarily chosen prior. T o retrieve a generative mechanism to sample new data points, we introduce an ex- post density estimation step that can be readily applied to the proposed framewo rk as well as existing VAEs, improving their sample quality. We show, in a rigor ous empirical study, that the proposed regularized deterministic autoencoders ar e able to generate samples that are comparable to, or better than, those of VAEs and more powerful alternatives when applied to images as well as to structured data such as molecules.

****

Adversarially Robust Representations with Smooth Encoders

Taylan Cemgil,Sumedh Ghaisas,Krishnamurthy (Dj) Dvijotham,Pushmeet Kohli

This paper studies the undesired phenomena of over-sensitivity of representation s learned by deep networks to semantically-irrelevant changes in data. We identi fy a cause for this shortcoming in the classical Variational Auto-encoder (VAE) objective, the evidence lower bound (ELBO). We show that the ELBO fails to contr ol the behaviour of the encoder out of the support of the empirical data distrib ution and this behaviour of the VAE can lead to extreme errors in the learned re presentation. This is a key hurdle in the effective use of representations for d ata-efficient learning and transfer. To address this problem, we propose to augm ent the data with specifications that enforce insensitivity of the representatio n with respect to families of transformations. To incorporate these specificatio ns, we propose a regularization method that is based on a selection mechanism th at creates a fictive data point by explicitly perturbing an observed true data p oint. For certain choices of parameters, our formulation naturally leads to the minimization of the entropy regularized Wasserstein distance between representat ions. We illustrate our approach on standard datasets and experimentally show th at significant improvements in the downstream adversarial accuracy can be achiev

ed by learning robust representations completely in an unsupervised manner, with out a reference to a particular downstream task and without a costly supervised adversarial training procedure.

****
AssembleNet: Searching for Multi-Stream Neural Connectivity in Video Architectures
Michael S. Ryoo,AJ Piergiovanni,Mingxing Tan,Anelia Angelova
Learning to represent videos is a very challenging task both algorithmically and computationally. Standard video CNN architectures have been designed by directly extending architectures devised for image understanding to include the time dimension, using modules such as 3D convolutions, or by using two-stream design to capture both appearance and motion in videos. We interpret a video CNN as a collection of multi-stream convolutional blocks connected to each other, and propose the approach of automatically finding neural architectures with better connectivity and spatio-temporal interactions for video understanding. This is done by evolving a population of overly-connected architectures guided by connection weight learning.
Architectures combining representations that abstract different input types (i.e., RGB and optical flow) at multiple temporal resolutions are searched for, allowing different types or sources of information to interact with each other. Our method, referred to as AssembleNet, outperforms prior approaches on public video datasets, in some cases by a great margin. We obtain 58.6% mAP on Charades and 34.27% accuracy on Moments-in-Time.
****
Representation Quality Explain Adversarial Attacks
Danilo Vasconcellos Vargas,Shashank Kotyan,Moe Matsuki
Neural networks have been shown vulnerable to adversarial samples. Slightly perturbed input images are able to change the classification of accurate models, showing that the representation learned is not as good as previously thought. To aid the development of better neural networks, it would be important to evaluate to what extent are current neural networks' representations capturing the existing features. Here we propose a way to evaluate the representation quality of neural networks using a novel type of zero-shot test, entitled Raw Zero-Shot. The main idea lies in the fact that some features are present on unknown classes and that unknown classes can be defined as a combination of previous learned features without representation bias (a bias towards representation that maps only current set of input-outputs and their boundary). To evaluate the soft-labels of unknown classes, two metrics are proposed. One is based on clustering validation techniques (Davies-Bouldin Index) and the other is based on soft-label distance of a given correct soft-label.
Experiments show that such metrics are in accordance with the robustness to adversarial attacks and might serve as a guidance to build better models as well as be used in loss functions to create new types of neural networks. Interestingly, the results suggests that dynamic routing networks such as CapsNet have better representation while current deeper DNNs are trading off representation quality for accuracy.
****
Inferring Dynamical Systems with Long-Range Dependencies through Line Attractor Regularization
Dominik Schmidt,Georgia Koppe,Max Beutelspacher,Daniel Durstewitz
Vanilla RNN with ReLU activation have a simple structure that is amenable to systematic dynamical systems analysis and interpretation, but they suffer from the exploding vs. vanishing gradients problem. Recent attempts to retain this simplicity while alleviating the gradient problem are based on proper initialization schemes or orthogonality/unitary constraints on the RNN’s recurrency matrix, which, however, comes with limitations to its expressive power with regards to dynamical systems phenomena like chaos or multi-stability. Here, we instead suggest a regularization scheme that pushes part of the RNN’s latent subspace toward a line attractor configuration that enables long short-term memory and arbitrarily s

low time scales. We show that our approach excels on a number of benchmarks like the sequential MNIST or multiplication problems, and enables reconstruction of dynamical systems which harbor widely different time scales.
****

End-To-End Input Selection for Deep Neural Networks
Stefan Oehmcke,Fabian Gieseke
Data have often to be moved between servers and clients during the inference phase. This is the case, for instance, when large amounts of data are stored on a public storage server without the possibility for the users to directly execute code and, hence, apply machine learning models. Depending on the available bandwidth, this data transfer can become a major bottleneck. We propose a simple yet effective framework that allows to select certain parts of the input data needed for the subsequent application of a given neural network. Both the associated selection masks as well as the neural network are trained simultaneously such that a good model performance is achieved while, at the same time, only a minimal amount of data is selected. During the inference phase, only the parts selected by the masks have to be transferred between the server and the client. Our experiments indicate that it is often possible to significantly reduce the amount of data needed to be transferred without affecting the model performance much.
****

Hierarchical Graph-to-Graph Translation for Molecules
Wengong Jin,Regina Barzilay,Tommi Jaakkola
The problem of accelerating drug discovery relies heavily on automatic tools to optimize precursor molecules to afford them with better biochemical properties. Our work in this paper substantially extends prior state-of-the-art on graph-to-graph translation methods for molecular optimization. In particular, we realize coherent multi-resolution representations by interweaving the encoding of substructure components with the atom-level encoding of the original molecular graph. Moreover, our graph decoder is fully autoregressive, and interleaves each step of adding a new substructure with the process of resolving its attachment to the emerging molecule. We evaluate our model on multiple molecular optimization tasks and show that our model significantly outperforms previous state-of-the-art baselines.
****

ReClor: A Reading Comprehension Dataset Requiring Logical Reasoning
Weihao Yu,Zihang Jiang,Yanfei Dong,Jiashi Feng
Recent powerful pre-trained language models have achieved remarkable performance on most of the popular datasets for reading comprehension. It is time to introduce more challenging datasets to push the development of this field towards more comprehensive reasoning of text. In this paper, we introduce a new Reading Comprehension dataset requiring logical reasoning (ReClor) extracted from standardized graduate admission examinations. As earlier studies suggest, human-annotated datasets usually contain biases, which are often exploited by models to achieve high accuracy without truly understanding the text. In order to comprehensively evaluate the logical reasoning ability of models on ReClor, we propose to identify biased data points and separate them into EASY set while the rest as HARD set. Empirical results show that state-of-the-art models have an outstanding ability to capture biases contained in the dataset with high accuracy on EASY set. However, they struggle on HARD set with poor performance near that of random guess, indicating more research is needed to essentially enhance the logical reasoning ability of current models.
****

DeepEnFM: Deep neural networks with Encoder enhanced Factorization Machine
Qiang Sun,Zhinan Cheng,Yanwei Fu,Wenxuan Wang,Yu-Gang Jiang,Xiangyang Xue
Click Through Rate (CTR) prediction is a critical task in industrial applications, especially for online social and commerce applications. It is challenging to find a proper way to automatically discover the effective cross features in CTR tasks. We propose a novel model for CTR tasks, called Deep neural networks with Encoder enhanced Factorization Machine (DeepEnFM). Instead of learning the cross features directly, DeepEnFM adopts the Transformer encoder as a backbone to ali

gn the feature embeddings with the clues of other fields. The embeddings generat ed from encoder are beneficial for the further feature interactions. Particularl y, DeepEnFM utilizes a bilinear approach to generate different similarity functi ons with respect to different field pairs. Furthermore, the max-pooling method m akes DeepEnFM feasible to capture both the supplementary and suppressing informa tion among different attention heads. Our model is validated on the Criteo and A vazu datasets, and achieves state-of-art performance.
****
A NEW POINTWISE CONVOLUTION IN DEEP NEURAL NETWORKS THROUGH EXTREMELY FAST AND N ON PARAMETRIC TRANSFORMS
Joonhyun Jeong,Sung-Ho Bae
Some conventional transforms such as Discrete Walsh-Hadamard Transform (DWHT) and Discrete Cosine Transform (DCT) have been widely used as feature extractor s in image processing but rarely applied in neural networks. However, we found t hat these conventional transforms have the ability to capture the cross-channel correlations without any learnable parameters in DNNs. This paper firstly propos es to apply conventional transforms on pointwise convolution, showing that such transforms significantly reduce the computational complexity of neural networks without accuracy performance degradation. Especially for DWHT, it requires no fl oating point multiplications but only additions and subtractions, which can cons iderably reduce computation overheads. In addition, its fast algorithm further r educes complexity of floating point addition from O(n^2) to O(nlog n). These non -parametric and low computational properties construct extremely efficient netwo rks in the number parameters and operations, enjoying accuracy gain. Our propose d DWHT-based model gained 1.49% accuracy increase with 79.4% reduced parameters and 48.4% reduced FLOPs compared with its baseline model (MoblieNet-V1) on the C IFAR 100 dataset.
****
Decaying momentum helps neural network training
John Chen,Anastasios Kyrillidis
Momentum is a simple and popular technique in deep learning for gradient-based o ptimizers. We propose a decaying momentum (Demon) rule, motivated by decaying th e total contribution of a gradient to all future updates. Applying Demon to Adam leads to significantly improved training, notably competitive to momentum SGD w ith learning rate decay, even in settings in which adaptive methods are typicall y non-competitive. Similarly, applying Demon to momentum SGD rivals momentum SGD with learning rate decay, and in many cases leads to improved performance. Demo n is trivial to implement and incurs limited extra computational overhead, compa red to the vanilla counterparts.
****
Regularizing Black-box Models for Improved Interpretability
Gregory Plumb,Maruan Al-Shedivat,Eric Xing,Ameet Talwalkar
Most of the work on interpretable machine learning has focused on designingeithe r inherently interpretable models, which typically trade-off accuracyfor interpr etability, or post-hoc explanation systems, which lack guarantees about their ex planation quality. We explore an alternative to theseapproaches by directly reg ularizing a black-box model for interpretabilityat training time. Our approach explicitly connects three key aspects ofinterpretable machine learning: (i) the model’s internal interpretability, (ii)the explanation system used at test time , and (iii) the metrics that measureexplanation quality. Our regularization res ults in substantial improvementin terms of the explanation fidelity and stabilit y metrics across a range ofdatasets and black-box explanation systems while slig htly improving accuracy. Finally, we justify theoretically that the benefits of our regularizationgeneralize to unseen points.
****
GPNET: MONOCULAR 3D VEHICLE DETECTION BASED ON LIGHTWEIGHT WHEEL GROUNDING POINT DETECTION NETWORK
zizhang.wu
We present a method to infer 3D location and orientation of vehicles on a single image. To tackle this problem, we optimize the mapping relation between the veh

icle’s wheel grounding point on the image and the real location of the wheel in the 3D real world coordinate. Here we also integrate three task priors, includin g a ground plane constraint and vehicle wheel grounding point position, as well as a small projection error from the image to the ground plane. And a robust lig ht network for grounding point detection in autopilot is proposed based on the v ehicle and wheel detection result. In the light grounding point detection networ k, the DSNT key point regression method is used for balancing the speed of conve rgence and the accuracy of position, which has been proved more robust and accur ate compared with the other key point detection methods. With more, the size of grounding point detection network is less than 1 MB, which can be executed quick ly on the embedded environment. The code will be available soon.
****

Needles in Haystacks: On Classifying Tiny Objects in Large Images
Nick Pawlowski,Suvrat Bhooshan,Nicolas Ballas,Francesco Ciompi,Ben Glocker,Micha l Drozdzal
In some important computer vision domains, such as medical or hyperspectral imag ing, we care about the classification of tiny objects in large images. However, most Convolutional Neural Networks (CNNs) for image classification were develope d using biased datasets that contain large objects, in mostly central image posi tions. To assess whether classical CNN architectures work well for tiny object c lassification we build a comprehensive testbed containing two datasets: one deri ved from MNIST digits and one from histopathology images. This testbed allows co ntrolled experiments to stress-test CNN architectures with a broad spectrum of s ignal-to-noise ratios. Our observations indicate that: (1) There exists a limit to signal-to-noise below which CNNs fail to generalize and that this limit is af fected by dataset size - more data leading to better performances; however, the amount of training data required for the model to generalize scales rapidly with the inverse of the object-to-image ratio (2) in general, higher capacity models exhibit better generalization; (3) when knowing the approximate object sizes, a dapting receptive field is beneficial; and (4) for very small signal-to-noise ra tio the choice of global pooling operation affects optimization, whereas for rel atively large signal-to-noise values, all tested global pooling operations exhib it similar performance.
****

The advantage of using Student's t-priors in variational autoencoders
Najmeh Abiri,Mattias Ohlsson
Is it optimal to use the standard Gaussian prior in variational autoencoders? Wi th Gaussian distributions, which are not weakly informative priors, variational autoencoders struggle to reconstruct the actual data. We provide numerical evide nce that encourages using Student's t-distributions as default priors in variati onal autoencoders, and we challenge the usual setup for the variational autoenco der structure by comparing Gaussian and Student's t-distribution priors with dif ferent forms of the covariance matrix.
****

Finite Depth and Width Corrections to the Neural Tangent Kernel
Boris Hanin,Mihai Nica
We prove the precise scaling, at finite depth and width, for the mean and varian ce of the neural tangent kernel (NTK) in a randomly initialized ReLU network. Th e standard deviation is exponential in the ratio of network depth to width. Thus , even in the limit of infinite overparameterization, the NTK is not determinist ic if depth and width simultaneously tend to infinity. Moreover, we prove that f or such deep and wide networks, the NTK has a non-trivial evolution during train ing by showing that the mean of its first SGD update is also exponential in the ratio of network depth to width. This is sharp contrast to the regime where dept h is fixed and network width is very large. Our results suggest that, unlike rel atively shallow and wide networks, deep and wide ReLU networks are capable of le arning data-dependent features even in the so-called lazy training regime.
****

Order Learning and Its Application to Age Estimation
Kyungsun Lim,Nyeong-Ho Shin,Young-Yoon Lee,Chang-Su Kim

We propose order learning to determine the order graph of classes, representing ranks or priorities, and classify an object instance into one of the classes. To this end, we design a pairwise comparator to categorize the relationship between two instances into one of three cases: one instance is `greater than,' `similar to,' or `smaller than' the other. Then, by comparing an input instance with reference instances and maximizing the consistency among the comparison results, the class of the input can be estimated reliably. We apply order learning to develop a facial age estimator, which provides the state-of-the-art performance. Moreover, the performance is further improved when the order graph is divided into disjoint chains using gender and ethnic group information or even in an unsupervised manner.

****

Model-based Saliency for the Detection of Adversarial Examples

Lisa Schut,Yarin Gal

Adversarial perturbations cause a shift in the salient features of an image, which may result in a misclassification. We demonstrate that gradient-based saliency approaches are unable to capture this shift, and develop a new defense which detects adversarial examples based on learnt saliency models instead. We study two approaches: a CNN trained to distinguish between natural and adversarial images using the saliency masks produced by our learnt saliency model, and a CNN trained on the salient pixels themselves as its input. On MNIST, CIFAR-10 and ASSIRA, our defenses are able to detect various adversarial attacks, including strong attacks such as C&W and DeepFool, contrary to gradient-based saliency and detectors which rely on the input image. The latter are unable to detect adversarial images when the L_2- and L_infinity- norms of the perturbations are too small. Lastly, we find that the salient pixel based detector improves on saliency map based detectors as it is more robust to white-box attacks.

****

Online Meta-Critic Learning for Off-Policy Actor-Critic Methods

Wei Zhou,Yiying Li,Yongxin Yang,Huaimin Wang,Timothy M. Hospedales

Off-Policy Actor-Critic (Off-PAC) methods have proven successful in a variety of continuous control tasks. Normally, the critic’s action-value function is updated using temporal-difference, and the critic in turn provides a loss for the actor that trains it to take actions with higher expected return. In this paper, we introduce a novel and flexible meta-critic that observes the learning process and meta-learns an additional loss for the actor that accelerates and improves actor-critic learning. Compared to the vanilla critic, the meta-critic network is explicitly trained to accelerate the learning process; and compared to existing meta-learning algorithms, meta-critic is rapidly learned online for a single task, rather than slowly over a family of tasks. Crucially, our meta-critic framework is designed for off-policy based learners, which currently provide state-of-the-art reinforcement learning sample efficiency. We demonstrate that online meta-critic learning leads to improvements in a variety of continuous control environments when combined with contemporary Off-PAC methods DDPG, TD3 and the state-of-the-art SAC.

****

BUZz: BUffer Zones for defending adversarial examples in image classification

Phuong Ha Nguyen*,Kaleel Mahmood*,Lam M. Nguyen,Thanh Nguyen,Marten van Dijk

We propose a novel defense against all existing gradient based adversarial attacks on deep neural networks for image classification problems. Our defense is based on a combination of deep neural networks and simple image transformations. While straight forward in implementation, this defense yields a unique security property which we term buffer zones. In this paper, we formalize the concept of buffer zones. We argue that our defense based on buffer zones is secure against state-of-the-art black box attacks. We are able to achieve this security even when the adversary has access to the entire original training data set and unlimited query access to the defense. We verify our security claims through experimentation using FashionMNIST, CIFAR-10 and CIFAR-100. We demonstrate <10% attack success rate -- significantly lower than what other well-known defenses offer -- at only a price of a 15-20% drop in clean accuracy. By using a new intuitive metric

we explain why this trade-off offers a significant improvement over prior work.
****
Efficient and Information-Preserving Future Frame Prediction and Beyond
Wei Yu,Yichao Lu,Steve Easterbrook,Sanja Fidler
Applying resolution-preserving blocks is a common practice to maximize information preservation in video prediction, yet their high memory consumption greatly limits their application scenarios. We propose CrevNet, a Conditionally Reversible Network that uses reversible architectures to build a bijective two-way autoencoder and its complementary recurrent predictor. Our model enjoys the theoretically guaranteed property of no information loss during the feature extraction, much lower memory consumption and computational efficiency. The lightweight nature of our model enables us to incorporate 3D convolutions without concern of memory bottleneck, enhancing the model's ability to capture both short-term and long-term temporal dependencies. Our proposed approach achieves state-of-the-art results on Moving MNIST, Traffic4cast and KITTI datasets. We further demonstrate the transferability of our self-supervised learning method by exploiting its learnt features for object detection on KITTI. Our competitive results indicate the potential of using CrevNet as a generative pre-training strategy to guide downstream tasks.
****
Path Space for Recurrent Neural Networks with ReLU Activations
Yue Wang,Qi Meng,Wei Chen,Yuting Liu,Zhi-Ming Ma,Tie-Yan Liu
It is well known that neural networks with rectified linear units (ReLU) activation functions are positively scale-invariant (i.e., the neural network is invariant to positive rescaling of weights). Optimization algorithms like stochastic gradient descent that optimize the neural networks in the vector space of weights, which are not positively scale-invariant. To solve this mismatch, a new parameter space called path space has been proposed for feedforward and convolutional neural networks. The path space is positively scale-invariant and optimization algorithms operating in path space have been shown to be superior than that in the original weight space. However, the theory of path space and the corresponding optimization algorithm cannot be naturally extended to more complex neural networks, like Recurrent Neural Networks(RNN) due to the recurrent structure and the parameter sharing scheme over time. In this work, we aim to construct path space for RNN with ReLU activations so that we can employ optimization algorithms in path space. To achieve the goal, we propose leveraging the reduction graph of RNN which removes the influence of time-steps, and prove that all the values of whose paths can serve as a sufficient representation of the RNN with ReLU activations. We then prove that the path space for RNN is composed by the basis paths in reduction graph, and design a \emph{Skeleton Method} to identify the basis paths efficiently. With the identified basis paths, we develop the optimization algorithm in path space for RNN models. Our experiments on several benchmark datasets show that we can obtain significantly more effective RNN models in this way than using optimization methods in the weight space.
****
Wasserstein Adversarial Regularization (WAR) on label noise
Bharath Damodaran,Kilian Fatras,Sylvain Lobry,Rémi Flamary,Devis Tuia,Nicolas Courty
Noisy labels often occur in vision datasets, especially when they are obtained from crowdsourcing or Web scraping. We propose a new regularization method, which enables learning robust classifiers in presence of noisy data. To achieve this goal, we propose a new adversarial regularization scheme based on the Wasserstein distance. Using this distance allows taking into account specific relations between classes by leveraging the geometric properties of the labels space. Our Wasserstein Adversarial Regularization (WAR) encodes a selective regularization, which promotes smoothness of the classifier between some classes, while preserving sufficient complexity of the decision boundary between others. We first discuss how and why adversarial regularization can be used in the context of label noise and then show the effectiveness of our method on five datasets corrupted with noisy labels: in both benchmarks and real datasets, WAR outperforms the state

-of-the-art
competitors.
****
Self-Supervised Speech Recognition via Local Prior Matching
Wei-Ning Hsu,Ann Lee,Gabriel Synnaeve,Awni Hannun
We propose local prior matching (LPM), a self-supervised objective for speech re cognition. The LPM objective leverages a strong language model to provide learni ng signal given unlabeled speech. Since LPM uses a language model, it can take a dvantage of vast quantities of both unpaired text and speech. The loss is theore tically well-motivated and simple to implement. More importantly, LPM is effecti ve. Starting from a model trained on 100 hours of labeled speech, with an additi onal 360 hours of unlabeled data LPM reduces the WER by 26% and 31% relative on a clean and noisy test set, respectively. This bridges the gap by 54% and 73% WE R on the two test sets relative to a fully supervised model on the same 360 hour s with labels. By augmenting LPM with an additional 500 hours of noisy data, we further improve the WER on the noisy test set by 15% relative. Furthermore, we p erform extensive ablative studies to show the importance of various configuratio ns of our self-supervised approach.
****
SRDGAN: learning the noise prior for Super Resolution with Dual Generative Adver sarial Networks
Jingwei GUAN,Cheng PAN,Songnan LI and Dahai YU
Single Image Super Resolution (SISR) is the task of producing a high resolution (HR) image from a given low-resolution (LR) image. It is a well researched prob- lem with extensive commercial applications like digital camera, video compres- sion, medical imaging, etc. Most recent super resolution works focus on the fea- ture learning architecture, like Chao Dong (2016); Dong et al. (2016); Wang et al. (2018b); Ledig et al. (2017). However, these works suffer from the following chal- lenges: (1) The low-resolution (LR) training images are artificially synt hesized us- ing HR images with bicubic downsampling, which have much more inform ation than real demosaic-upscaled images. The mismatch between training and real istic mobile data heavily blocks the effect on practical SR problem. (2) These m ethods cannot effectively handle the blind distortions during super resolution i n practical applications. In this work, an end-to-end novel framework, including high-to-low network and low-to-high network, is proposed to solve the above pro blems with dual Generative Adversarial Networks (GAN). First, the above mismatch prob- lems are well explored with the high-to-low network, where clear high-res olution image and the corresponding realistic low-resolution image pairs can be gener- ated. With high-to-low network, a large-scale General Mobile Super Resolu tion Dataset, GMSR, is proposed, which can be utilized for training or as a benc h- mark for super resolution methods. Second, an effective low-to-high network (super resolution network) is proposed in the framework. Benefiting from the GMSR dataset and novel training strategies, the proposed super resolution model can effectively handle detail recovery and denoising at the same time.
****
Amata: An Annealing Mechanism for Adversarial Training Acceleration
Nanyang Ye,Qianxiao Li,Zhanxing Zhu
Despite of the empirical success in various domains, it has been revealed that d eep neural networks are vulnerable to maliciously perturbed input data that much degrade their performance. This is known as adversarial attacks. To counter adv ersarial attacks, adversarial training formulated as a form of robust optimizati on has been demonstrated to be effective. However, conducting adversarial traini ng brings much computational overhead compared with standard training. In order to reduce the computational cost, we propose a simple yet effective modification to the commonly used projected gradient descent (PGD) adversarial training by i ncreasing the number of adversarial training steps and decreasing the adversaria l training step size gradually as training proceeds. We analyze the optimality o f this annealing mechanism through the lens of optimal control theory, and we al so prove the convergence of our proposed algorithm. Numerical experiments on sta ndard datasets, such as MNIST and CIFAR10, show that our method can achieve simi

lar or even better robustness with around 1/3 to 1/2 computation time compared w
ith PGD.
****

Context Based Machine Translation With Recurrent Neural Network For English-Amha
ric Translation
Yeabsira Asefa Ashengo,Rosa Tsegaye Aga,Surafel Lemma Abebe
The current approaches for machine translation usually require large set of para
llel corpus in order to achieve fluency like in the case of neural machine trans
lation (NMT), statistical machine translation (SMT) and example-based machine tr
anslation (EBMT). The context awareness of phrase-based machine translation (PBM
T) approaches is also questionable. This research develops a system that transla
tes English text to Amharic text using a combination of context based machine tr
anslation (CBMT) and a recurrent neural network machine translation (RNNMT). We
built a bilingual dictionary for the CBMT system to use along with a large targe
t corpus. The RNNMT model has then been provided with the output of the CBMT and
 a parallel corpus for training. Our combinational approach on English-Amharic l
anguage pair yields a performance improvement over the simple neural machine tra
nslation (NMT).
****

Robust Domain Randomization for Reinforcement Learning
Reda Bahi Slaoui,William R. Clements,Jakob N. Foerster,Sébastien Toth
Producing agents that can generalize to a wide range of environments is a signif
icant challenge in reinforcement learning. One method for overcoming this issue
is domain randomization, whereby at the start of each training episode some para
meters of the environment are randomized so that the agent is exposed to many po
ssible variations. However, domain randomization is highly inefficient and may l
ead to policies with high variance across domains. In this work, we formalize th
e domain randomization problem, and show that minimizing the policy's Lipschitz
constant with respect to the randomization parameters leads to low variance in t
he learned policies. We propose a method where the agent only needs to be traine
d on one variation of the environment, and its learned state representations are
 regularized during training to minimize this constant. We conduct experiments t
hat demonstrate that our technique leads to more efficient and robust learning t
han standard domain randomization, while achieving equal generalization scores.
****

NAS evaluation is frustratingly hard
Antoine Yang,Pedro M. Esperança,Fabio M. Carlucci
Neural Architecture Search (NAS) is an exciting new field which promises to be a
s much as a game-changer as Convolutional Neural Networks were in 2012. Despite
many great works leading to substantial improvements on a variety of tasks, comp
arison between different methods is still very much an open issue. While most al
gorithms are tested on the same datasets, there is no shared experimental protoc
ol followed by all. As such, and due to the under-use of ablation studies, there
 is a lack of clarity regarding why certain methods are more effective than othe
rs. Our first contribution is a benchmark of 8 NAS methods on 5 datasets. To ove
rcome the hurdle of comparing methods with different search spaces, we propose u
sing a method’s relative improvement over the randomly sampled average architect
ure, which effectively removes advantages arising from expertly engineered searc
h spaces or training protocols. Surprisingly, we find that many NAS techniques s
truggle to significantly beat the average architecture baseline. We perform furt
her experiments with the commonly used DARTS search space in order to understand
 the contribution of each component in the NAS pipeline. These experiments highl
ight that: (i) the use of tricks in the evaluation protocol has a predominant im
pact on the reported performance of architectures; (ii) the cell-based search sp
ace has a very narrow accuracy range, such that the seed has a considerable impa
ct on architecture rankings; (iii) the hand-designed macrostructure (cells) is m
ore important than the searched micro-structure (operations); and (iv) the depth
-gap is a real phenomenon, evidenced by the change in rankings between 8 and 20
cell architectures. To conclude, we suggest best practices, that we hope will pr
ove useful for the community and help mitigate current NAS pitfalls, e.g. diffic

ulties in reproducibility and comparison of search methods. The
code used is available at https://github.com/antoyang/NAS-Benchmark.
****
Ellipsoidal Trust Region Methods for Neural Network Training
Leonard Adolphs,Jonas Kohler,Aurelien Lucchi
We investigate the use of ellipsoidal trust region constraints for second-order optimization of neural networks. This approach can be seen as a higher-order counterpart of adaptive gradient methods, which we here show to be interpretable as first-order trust region methods with ellipsoidal constraints. In particular, we show that the preconditioning matrix used in RMSProp and Adam satisfies the necessary conditions for provable convergence of second-order trust region methods with standard worst-case complexities. Furthermore, we run experiments across different neural architectures and datasets to find that the ellipsoidal constraints constantly outperform their spherical counterpart both in terms of number of backpropagations and asymptotic loss value. Finally, we find comparable performance to state-of-the-art first-order methods in terms of backpropagations, but further advances in hardware are needed to render Newton methods competitive in terms of time.
****
Learning Semantically Meaningful Representations Through Embodiment
Viviane Clay,Peter König,Kai-Uwe Kühnberger,Gordon Pipa
How do humans acquire a meaningful understanding of the world with little to no supervision or semantic labels provided by the environment? Here we investigate embodiment and a closed loop between action and perception as one key component in this process. We take a close look at the representations learned by a deep reinforcement learning agent that is trained with visual and vector observations collected in a 3D environment with sparse rewards. We show that this agent learns semantically meaningful and stable representations of its environment without receiving any semantic labels. Our results show that the agent learns to represent the action relevant information extracted from pixel input in a wide variety of sparse activation patterns. The quality of the representations learned shows the strength of embodied learning and its advantages over fully supervised approaches with regards to robustness and generalizability.
****
Superseding Model Scaling by Penalizing Dead Units and Points with Separation Constraints
Carles Riera,Camilo Rey-Torres,Eloi Puertas,Oriol Pujol
In this article, we study a proposal that enables to train extremely thin (4 or 8 neurons per layer) and relatively deep (more than 100 layers) feedforward networks without resorting to any architectural modification such as Residual or Dense connections, data normalization or model scaling. We accomplish that by alleviating two problems. One of them are neurons whose output is zero for all the dataset, which renders them useless. This problem is known to the academic community as \emph{dead neurons}. The other is a less studied problem, dead points. Dead points refers to data points that are mapped to zero during the forward pass of the network. As such, the gradient generated by those points is not propagated back past the layer where they die, thus having no effect in the training process. In this work, we characterize both problems and propose a constraint formulation that added to the standard loss function solves them both. As an additional benefit, the proposed method allows to initialize the network weights with constant or even zero values and still allowing the network to converge to reasonable results. We show very promising results on a toy, MNIST, and CIFAR-10 datasets.
****
Robust Graph Representation Learning via Neural Sparsification
Cheng Zheng,Bo Zong,Wei Cheng,Dongjin Song,Jingchao Ni,Wenchao Yu,Haifeng Chen,Wei Wang
Graph representation learning serves as the core of many important prediction tasks, ranging from product recommendation in online marketing to fraud detection in financial domain. Real-life graphs are usually large with complex local neigh

borhood, where each node is described by a rich set of features and easily conne cts to dozens or even hundreds of neighbors. Most existing graph learning techni ques rely on neighborhood aggregation, however, the complexity on real-life grap hs is usually high, posing non-trivial overfitting risk during model training. I n this paper, we present Neural Sparsification (NeuralSparse), a supervised grap h sparsification technique that mitigates the overfitting risk by reducing the c omplexity of input graphs. Our method takes both structural and non-structural i nformation as input, utilizes deep neural networks to parameterize the sparsific ation process, and optimizes the parameters by feedback signals from downstream tasks. Under the NeuralSparse framework, supervised graph sparsification could s eamlessly connect with existing graph neural networks for more robust performanc e on testing data. Experimental results on both benchmark and private datasets s how that NeuralSparse can effectively improve testing accuracy and bring up to 7 .4% improvement when working with existing graph neural networks on node classif ication tasks.

****

Hyperbolic Discounting and Learning Over Multiple Horizons

William Fedus,Carles Gelada,Yoshua Bengio,Marc G. Bellemare,Hugo Larochelle

Reinforcement learning (RL) typically defines a discount factor as part of the M arkov Decision Process. The discount factor values future rewards by an exponen tial scheme that leads to theoretical convergence guarantees of the Bellman equa tion. However, evidence from psychology, economics and neuroscience suggests tha t humans and animals instead have hyperbolic time-preferences. Here we extend e arlier work of Kurth-Nelson and Redish and propose an efficient deep reinforceme nt learning agent that acts via hyperbolic discounting and other non-exponential discount mechanisms. We demonstrate that a simple approach approximates hyperbo lic discount functions while still using familiar temporal-difference learning t echniques in RL. Additionally, and independent of hyperbolic discounting, we ma ke a surprising discovery that simultaneously learning value functions over mult iple time-horizons is an effective auxiliary task which often improves over stat e-of-the-art methods.

****

CLN2INV: Learning Loop Invariants with Continuous Logic Networks

Gabriel Ryan,Justin Wong,Jianan Yao,Ronghui Gu,Suman Jana

Program verification offers a framework for ensuring program correctness and the refore systematically eliminating different classes of bugs. Inferring loop inva riants is one of the main challenges behind automated verification of real-world programs which often contain many loops. In this paper, we present the Continuo us Logic Network (CLN), a novel neural architecture for automatically learning l oop invariants directly from program execution traces. Unlike existing neural ne tworks, CLNs can learn precise and explicit representations of formulas in Satis fiability Modulo Theories (SMT) for loop invariants from program execution trac es. We develop a new sound and complete semantic mapping for assigning SMT formu las to continuous truth values that allows CLNs to be trained efficiently. We us e CLNs to implement a new inference system for loop invariants, CLN2INV, that si gnificantly outperforms existing approaches on the popular Code2Inv dataset. CLN 2INV is the first tool to solve all 124 theoretically solvable problems in the C ode2Inv dataset. Moreover, CLN2INV takes only 1.1 second on average for each pro blem, which is 40 times faster than existing approaches. We further demonstrate that CLN2INV can even learn 12 significantly more complex loop invariants than t he ones required for the Code2Inv dataset.

****

Federated User Representation Learning

Duc Bui,Kshitiz Malik,Jack Goetz,Seungwhan Moon,Honglei Liu,Anuj Kumar,Kang G. S hin

Collaborative personalization, such as through learned user representations (emb eddings), can improve the prediction accuracy of neural-network-based models sig nificantly. We propose Federated User Representation Learning (FURL), a simple, scalable, privacy-preserving and resource-efficient way to utilize existing neur al personalization techniques in the Federated Learning (FL) setting. FURL divid

es model parameters into federated and private parameters. Private parameters, s uch as private user embeddings, are trained locally, but unlike federated parame ters, they are not transferred to or averaged on the server. We show theoretical ly that this parameter split does not affect training for most model personaliza tion approaches. Storing user embeddings locally not only preserves user privacy , but also improves memory locality of personalization compared to on-server tra ining. We evaluate FURL on two datasets, demonstrating a significant improvement in model quality with 8% and 51% performance increases, and approximately the s ame level of performance as centralized training with only 0% and 4% reductions. Furthermore, we show that user embeddings learned in FL and the centralized set ting have a very similar structure, indicating that FURL can learn collaborative ly through the shared parameters while preserving user privacy.
****

INSTANCE CROSS ENTROPY FOR DEEP METRIC LEARNING
Xinshao Wang,Elyor Kodirov,Yang Hua,Neil M. Robertson
Loss functions play a crucial role in deep metric learning thus a variety of the m have been proposed. Some supervise the learning process by pairwise or triplet wise similarity constraints while others take the advantage of structured simila rity information among multiple data points. In this work, we approach deep metr ic learning from a novel perspective. We propose instance cross entropy (ICE) wh ich measures the difference between an estimated instance-level matching distrib ution and its ground-truth one. ICE has three main appealing properties. Firstly , similar to categorical cross entropy (CCE), ICE has clear probabilistic interp retation and exploits structured semantic similarity information for learning su pervision. Secondly, ICE is scalable to infinite training data as it learns on m ini-batches iteratively and is independent of the training set size. Thirdly, mo tivated by our relative weight analysis, seamless sample reweighting is incorpor ated. It rescales samples’ gradients to control the differentiation degree over training examples instead of truncating them by sample mining. In addition to it s simplicity and intuitiveness, extensive experiments on three real-world benchm arks demonstrate the superiority of ICE.
****

Scalable Neural Methods for Reasoning With a Symbolic Knowledge Base
William W. Cohen,Haitian Sun,R. Alex Hofer,Matthew Siegler
We describe a novel way of representing a symbolic knowledge base (KB) called a sparse-matrix reified KB. This representation enables neural modules that are f ully differentiable, faithful to the original semantics of the KB, expressive en ough to model multi-hop inferences, and scalable enough to use with realisticall y large KBs. The sparse-matrix reified KB can be distributed across multiple GPU s, can scale to tens of millions of entities and facts, and is orders of magnitu de faster than naive sparse-matrix implementations. The reified KB enables very simple end-to-end architectures to obtain competitive performance on several be nchmarks representing two families of tasks: KB completion, and learning semanti c parsers from denotations.
****

Variational pSOM: Deep Probabilistic Clustering with Self-Organizing Maps
Laura Manduchi,Matthias Hüser,Gunnar Rätsch,Vincent Fortuin
Generating visualizations and interpretations from high-dimensional data is a common problem in many fields. Two key approaches for tackling this problem are clustering and representation learning. There are very performant deep clustering models on the one hand and interpretable representation learning tech niques,
often relying on latent topological structures such as self-organizing maps, on the other hand. However, current methods do not yet successfully combine these two approaches. We present a new deep architecture for probabilistic clust ering,
VarPSOM, and its extension to time series data, VarTPSOM, composed of VarPSOM modules connected by LSTM cells. We show that they achieve superior clustering performance compared to current deep clustering methods on static MNIST/Fashion-MNIST data as well as medical time series, while inducing an

interpretable representation. Moreover, on the medical time series, VarTPSOM successfully predicts future trajectories in the original data space.
****
Augmenting Self-attention with Persistent Memory
Sainbayar Sukhbaatar,Edouard Grave,Guillaume Lample,Herve Jegou,Armand Joulin
Transformer networks have lead to important progress in language modeling and machine translation. These models include two consecutive modules, a feed-forward layer and a self-attention layer. The latter allows the network to capture long term dependencies and are often regarded as the key ingredient in the success of Transformers. Building upon this intuition, we propose a new model that solely consists of attention layers. More precisely, we augment the self-attention layers with persistent memory vectors that play a similar role as the feed-forward layer. Thanks to these vectors, we can remove the feed-forward layer without degrading the performance of a transformer. Our evaluation shows the benefits brought by our model on standard character and word level language modeling benchmarks.
****
Information Plane Analysis of Deep Neural Networks via Matrix--Based Renyi's Entropy and Tensor Kernels
Kristoffer Wickstrøm,Sigurd Løkse,Michael Kampffmeyer,Shujian Yu,Jose Principe,Robert Jenssen
Analyzing deep neural networks (DNNs) via information plane (IP) theory has gained tremendous attention recently as a tool to gain insight into, among others, their generalization ability. However, it is by no means obvious how to estimate mutual information (MI) between each hidden layer and the input/desired output, to construct the IP. For instance, hidden layers with many neurons require MI estimators with robustness towards the high dimensionality associated with such layers. MI estimators should also be able to naturally handle convolutional layers, while at the same time being computationally tractable to scale to large networks. None of the existing IP methods to date have been able to study truly deep Convolutional Neural Networks (CNNs), such as the e.g.\ VGG-16. In this paper, we propose an IP analysis using the new matrix--based R\'enyi's entropy coupled with tensor kernels over convolutional layers, leveraging the power of kernel methods to represent properties of the probability distribution independently of the dimensionality of the data. The obtained results shed new light on the previous literature concerning small-scale DNNs, however using a completely new approach. Importantly, the new framework enables us to provide the first comprehensive IP analysis of contemporary large-scale DNNs and CNNs, investigating the different training phases and providing new insights into the training dynamics of large-scale neural networks.
****
Ridge Regression: Structure, Cross-Validation, and Sketching
Sifan Liu,Edgar Dobriban
We study the following three fundamental problems about ridge regression: (1) what is the structure of the estimator? (2) how to correctly use cross-validation to choose the regularization parameter? and (3) how to accelerate computation without losing too much accuracy? We consider the three problems in a unified large-data linear model. We give a precise representation of ridge regression as a covariance matrix-dependent linear combination of the true parameter and the noise.
We study the bias of K-fold cross-validation for choosing the regularization parameter, and propose a simple bias-correction. We analyze the accuracy of primal and dual sketching for ridge regression, showing they are surprisingly accurate. Our results are illustrated by simulations and by analyzing empirical data.
****
Hindsight Trust Region Policy Optimization
Hanbo Zhang,Site Bai,Xuguang Lan,Nanning Zheng
As reinforcement learning continues to drive machine intelligence beyond its conventional boundary, unsubstantial practices in sparse reward environment severely limit further applications in a broader range of advanced fields. Motivated by

the demand for an effective deep reinforcement learning algorithm that accommodates sparse reward environment, this paper presents Hindsight Trust Region Policy Optimization (HTRPO), a method that efficiently utilizes interactions in sparse reward conditions to optimize policies within trust region and, in the meantime, maintains learning stability. Firstly, we theoretically adapt the TRPO objective function, in the form of the expected return of the policy, to the distribution of hindsight data generated from the alternative goals. Then, we apply Monte Carlo with importance sampling to estimate KL-divergence between two policies, taking the hindsight data as input. Under the condition that the distributions are sufficiently close, the KL-divergence is approximated by another f-divergence. Such approximation results in the decrease of variance and alleviates the instability during policy update. Experimental results on both discrete and continuous benchmark tasks demonstrate that HTRPO converges significantly faster than previous policy gradient methods. It achieves effective performances and high data-efficiency for training policies in sparse reward environments.
****
Policy Optimization with Stochastic Mirror Descent
Long Yang,Gang Zheng,Zavier Zhang,Yu Zhang,Qian Zheng,Jun Wen,Gang Pana sample efficient policy gradient method with stochastic mirror descent.
Improving sample efficiency has been a longstanding goal in reinforcement learning.
In this paper, we propose the $\mathtt{VRMPO}$: a sample efficient policy gradient method with stochastic mirror descent.
A novel variance reduced policy gradient estimator is the key of $\mathtt{VRMPO}$ to improve sample efficiency.
Our $\mathtt{VRMPO}$ needs only $\mathcal{O}(\epsilon^{-3})$ sample trajectories to achieve an ϵ-approximate first-order stationary point,
which matches the best-known sample complexity.
We conduct extensive experiments to show our algorithm outperforms state-of-the-art policy gradient methods in various settings.
****
Graph convolutional networks for learning with few clean and many noisy labels
Ahmet Iscen,Giorgos Tolias,Yannis Avrithis,Ondrej Chum,Cordelia Schmid
In this work we consider the problem of learning a classifier from noisy labels when a few clean labeled examples are given. The structure of clean and noisy data is modeled by a graph per class and Graph Convolutional Networks (GCN) are used to predict class relevance of noisy examples. For each class, the GCN is treated as a binary classifier learning to discriminate clean from noisy examples using a weighted binary cross-entropy loss function, and then the GCN-inferred "clean" probability is exploited as a relevance measure. Each noisy example is weighted by its relevance when learning a classifier for the end task. We evaluate our method on an extended version of a few-shot learning problem, where the few clean examples of novel classes are supplemented with additional noisy data. Experimental results show that our GCN-based cleaning process significantly improves the classification accuracy over not cleaning the noisy data and standard few-shot classification where only few clean examples are used. The proposed GCN-based method outperforms the transductive approach (Douze et al., 2018) that is using the same additional data without labels.
****
A Constructive Prediction of the Generalization Error Across Scales
Jonathan S. Rosenfeld,Amir Rosenfeld,Yonatan Belinkov,Nir Shavit
The dependency of the generalization error of neural networks on model and dataset size is of critical importance both in practice and for understanding the theory of neural networks. Nevertheless, the functional form of this dependency remains elusive. In this work, we present a functional form which approximates well the generalization error in practice. Capitalizing on the successful concept of model scaling (e.g., width, depth), we are able to simultaneously construct such a form and specify the exact models which can attain it across model/data scales. Our construction follows insights obtained from observations conducted over a range of model/data scales, in various model types and datasets, in vision and

language tasks. We show that the form both fits the observations well across sc
ales, and provides accurate predictions from small- to large-scale models and da
ta.
****
MLModelScope: A Distributed Platform for ML Model Evaluation and Benchmarking at
Scale
Cheng Li,Abdul Dakkak,Jinjun Xiong,Wen-mei Hwu
Machine Learning (ML) and Deep Learning (DL) innovations are being introduced at
such a rapid pace that researchers are hard-pressed to analyze and study them.
The complicated procedures for evaluating innovations, along with the lack of st
andard and efficient ways of specifying and provisioning ML/DL evaluation, is a
major "pain point" for the community. This paper proposes MLModelScope, an open-
source, framework/hardware agnostic, extensible and customizable design that ena
bles repeatable, fair, and scalable model evaluation and benchmarking. We imple
ment the distributed design with support for all major frameworks and hardware,
and equip it with web, command-line, and library interfaces. To demonstrate MLMo
delScope's capabilities we perform parallel evaluation and show how subtle chang
es to model evaluation pipeline affects the accuracy and HW/SW stack choices aff
ect performance.
****
A Mention-Pair Model of Annotation with Nonparametric User Communities
Silviu Paun,Juntao Yu,Jon Chamberlain,Udo Kruschwitz,Massimo Poesio
The availability of large datasets is essential for progress in coreference and
other areas of NLP. Crowdsourcing has proven a viable alternative to expert anno
tation, offering similar quality for better scalability. However, crowdsourcing
require adjudication, and most models of annotation focus on classification task
s where the set of classes is predetermined. This restriction does not apply to
anaphoric annotation, where coders relate markables to coreference chains whose
number cannot be predefined. This gap was recently covered with the introduction
of a mention pair model of anaphoric annotation (MPA). In this work we extend M
PA to alleviate the effects of sparsity inherent in some crowdsourcing environme
nts. Specifically, we use a nonparametric partially pooled structure (based on a
stick breaking process), fitting jointly with the ability of the annotators hie
rarchical community profiles. The individual estimates can thus be improved usin
g information about the community when the data is scarce. We show, using a rece
ntly published large-scale crowdsourced anaphora dataset, that the proposed mode
l performs better than its unpooled counterpart in conditions of sparsity, and o
n par when enough observations are available. The model is thus more resilient t
o different crowdsourcing setups, and, further provides insights into the commun
ity of workers. The model is also flexible enough to be used in standard annotat
ion tasks for classification where it registers on par performance with the stat
e of the art.
****
An Inductive Bias for Distances: Neural Nets that Respect the Triangle Inequalit
y
Silviu Pitis,Harris Chan,Kiarash Jamali,Jimmy Ba
Distances are pervasive in machine learning. They serve as similarity measures,
loss functions, and learning targets; it is said that a good distance measure so
lves a task. When defining distances, the triangle inequality has proven to be a
useful constraint, both theoretically---to prove convergence and optimality gua
rantees---and empirically---as an inductive bias. Deep metric learning architect
ures that respect the triangle inequality rely, almost exclusively, on Euclidean
distance in the latent space. Though effective, this fails to model two broad c
lasses of subadditive distances, common in graphs and reinforcement learning: as
ymmetric metrics, and metrics that cannot be embedded into Euclidean space. To a
ddress these problems, we introduce novel architectures that are guaranteed to s
atisfy the triangle inequality. We prove our architectures universally approxima
te norm-induced metrics on $\mathbb{R}^n$, and present a similar result for modi
fied Input Convex Neural Networks. We show that our architectures outperform exi
sting metric approaches when modeling graph distances and have a better inductiv

e bias than non-metric approaches when training data is limited in the multi-goa
l reinforcement learning setting.

****
NPTC-net: Narrow-Band Parallel Transport Convolutional Neural Network on Point C
louds
Pengfei Jin,Tianhao Lai,Rongjie Lai,Bin Dong
Convolution plays a crucial role in various applications in signal and image pro
cessing, analysis and recognition. It is also the main building block of convolu
tion neural networks (CNNs). Designing appropriate convolution neural networks o
n manifold-structured point clouds can inherit and empower recent advances of CN
Ns to analyzing and processing point cloud data. However, one of the major chall
enges is to define a proper way to "sweep" filters through the point cloud as a
natural generalization of the planar convolution and to reflect the point cloud'
s geometry at the same time. In this paper, we consider generalizing convolution
 by adapting parallel transport on the point cloud. Inspired by a triangulated s
urface based method \cite{DBLP:journals/corr/abs-1805-07857}, we propose the Nar
row-Band Parallel Transport Convolution (NPTC) using a specifically defined conn
ection on a voxelized narrow-band approximation of point cloud data. With that,
we further propose a deep convolutional neural network based on NPTC (called NPT
C-net) for point cloud classification and segmentation. Comprehensive experiment
s show that the proposed NPTC-net achieves similar or better results than curren
t state-of-the-art methods on point clouds classification and segmentation.
****
Mogrifier LSTM
Gábor Melis,Tomáš Ko■iský,Phil Blunsom
Many advances in Natural Language Processing have been based upon more expressiv
e models for how inputs interact with the context in which they occur. Recurrent
 networks, which have enjoyed a modicum of success, still lack the generalizatio
n and systematicity ultimately required for modelling language. In this work, we
 propose an extension to the venerable Long Short-Term Memory in the form of mut
ual gating of the current input and the previous output. This mechanism affords
the modelling of a richer space of interactions between inputs and their context
. Equivalently, our model can be viewed as making the transition function given
by the LSTM context-dependent. Experiments demonstrate markedly improved general
ization on language modelling in the range of 3–4 perplexity points on Penn Tree
bank and Wikitext-2, and 0.01–0.05 bpc on four character-based datasets. We esta
blish a new state of the art on all datasets with the exception of Enwik8, where
 we close a large gap between the LSTM and Transformer models.

****
Individualised Dose-Response Estimation using Generative Adversarial Nets
Ioana Bica,James Jordon,Mihaela van der Schaar
The problem of estimating treatment responses from observational data is by now
a well-studied one. Less well studied, though, is the problem of treatment respo
nse estimation when the treatments are accompanied by a continuous dosage parame
ter. In this paper, we tackle this lesser studied problem by building on a modif
ication of the generative adversarial networks (GANs) framework that has already
 demonstrated effectiveness in the former problem. Our model, DRGAN, is flexible
, capable of handling multiple treatments each accompanied by a dosage parameter
. The key idea is to use a significantly modified GAN model to generate entire d
ose-response curves for each sample in the training data which will then allow u
s to use standard supervised methods to learn an inference model capable of esti
mating these curves for a new sample. Our model consists of 3 blocks: (1) a gene
rator, (2) a discriminator, (3) an inference block. In order to address the chal
lenge presented by the introduction of dosages, we propose novel architectures f
or both our generator and discriminator. We model the generator as a multi-task
deep neural network. In order to address the increased complexity of the treatme
nt space (because of the addition of dosages), we develop a hierarchical discrim
inator consisting of several networks: (a) a treatment discriminator, (b) a dosa

ge discriminator for each treatment. In the experiments section, we introduce a new semi-synthetic data simulation for use in the dose-response setting and demonstrate improvements over the existing benchmark models.
****
Physics-as-Inverse-Graphics: Unsupervised Physical Parameter Estimation from Video
Miguel Jaques,Michael Burke,Timothy Hospedales
We propose a model that is able to perform physical parameter estimation of systems from video, where the differential equations governing the scene dynamics are known, but labeled states or objects are not available. Existing physical scene understanding methods require either object state supervision, or do not integrate with differentiable physics to learn interpretable system parameters and states. We address this problem through a \textit{physics-as-inverse-graphics} approach that brings together vision-as-inverse-graphics and differentiable physics engines, where objects and explicit state and velocity representations are discovered by the model. This framework allows us to perform long term extrapolative video prediction, as well as vision-based model-predictive control. Our approach significantly outperforms related unsupervised methods in long-term future frame prediction of systems with interacting objects (such as ball-spring or 3-body gravitational systems), due to its ability to build dynamics into the model as an inductive bias. We further show the value of this tight vision-physics integration by demonstrating data-efficient learning of vision-actuated model-based control for a pendulum system. We also show that the controller's interpretability provides unique capabilities in goal-driven control and physical reasoning for zero-data adaptation.
****
Trajectory representation learning for Multi-Task NMRDPs planning
Firas JARBOUI,Vianney PERCHET,Roman EGGER
Expanding Non Markovian Reward Decision Processes (NMRDP) into Markov Decision Processes (MDP) enables the use of state of the art Reinforcement Learning (RL) techniques to identify optimal policies. In this paper an approach to exploring NMRDPs and expanding them into MDPs, without the prior knowledge of the reward structure, is proposed. The non Markovianity of the reward function is disentangled under the assumption that sets of similar and dissimilar trajectory batches can be sampled. More precisely, within the same batch, measuring the similarity between any couple of trajectories is permitted, although comparing trajectories from different batches is not possible. A modified version of the triplet loss is optimised to construct a representation of the trajectories under which rewards become Markovian.
****
Incorporating Horizontal Connections in Convolution by Spatial Shuffling
Ikki Kishida,Hideki Nakayama
Convolutional Neural Networks (CNNs) are composed of multiple convolution layers and show elegant performance in vision tasks.
The design of the regular convolution is based on the Receptive Field (RF) where the information within a specific region is processed.
In the view of the regular convolution's RF, the outputs of neurons in lower layers with smaller RF are bundled to create neurons in higher layers with larger RF.
As a result, the neurons in high layers are able to capture the global context even though the neurons in low layers only see the local information.
However, in lower layers of the biological brain, the information outside of the RF changes the properties of neurons.
In this work, we extend the regular convolution and propose spatially shuffled convolution (ss convolution).
In ss convolution, the regular convolution is able to use the information outside of its RF by spatial shuffling which is a simple and lightweight operation.
We perform experiments on CIFAR-10 and ImageNet-1k dataset, and show that ss convolution improves the classification performance across various CNNs.
****

Is Deep Reinforcement Learning Really Superhuman on Atari? Leveling the playing field
Marin Toromanoff,Emilie Wirbel,Fabien Moutarde
Consistent and reproducible evaluation of Deep Reinforcement Learning (DRL) is not straightforward. In the Arcade Learning Environment (ALE), small changes in environment parameters such as stochasticity or the maximum allowed play time can lead to very different performance. In this work, we discuss the difficulties of comparing different agents trained on ALE. In order to take a step further towards reproducible and comparable DRL, we introduce SABER, a Standardized Atari BEnchmark for general Reinforcement learning algorithms. Our methodology extends previous recommendations and contains a complete set of environment parameters as well as train and test procedures. We then use SABER to evaluate the current state of the art, Rainbow. Furthermore, we introduce a human world records baseline, and argue that previous claims of expert or superhuman performance of DRL might not be accurate. Finally, we propose Rainbow-IQN by extending Rainbow with Implicit Quantile Networks (IQN) leading to new state-of-the-art performance. Source code is available for reproducibility.
****
Counterfactuals uncover the modular structure of deep generative models
Michel Besserve,Arash Mehrjou,Rémy Sun,Bernhard Schölkopf
Deep generative models can emulate the perceptual properties of complex image datasets, providing a latent representation of the data. However, manipulating such representation to perform meaningful and controllable transformations in the data space remains challenging without some form of supervision. While previous work has focused on exploiting statistical independence to \textit{disentangle} latent factors, we argue that such requirement can be advantageously relaxed and propose instead a non-statistical framework that relies on identifying a modular organization of the network, based on counterfactual manipulations. Our experiments support that modularity between groups of channels is achieved to a certain degree on a variety of generative models. This allowed the design of targeted interventions on complex image datasets, opening the way to applications such as computationally efficient style transfer and the automated assessment of robustness to contextual changes in pattern recognition systems.
****
Pushing the bounds of dropout
Gábor Melis,Charles Blundell,Tomáš Ko■iský,Karl Moritz Hermann,Chris Dyer,Phil Blunsom
We push on the boundaries of our knowledge about dropout by showing theoretically that dropout training can be understood as performing MAP estimation concurrently for an entire family of conditional models whose objectives are themselves lower bounded by the original dropout objective. This discovery allows us to pick any model from this family after training, which leads to a substantial improvement on regularisation-heavy language modelling. The family includes models that compute a power mean over the sampled dropout masks, and their less stochastic subvariants with tighter and higher lower bounds than the fully stochastic dropout objective. The deterministic subvariant's bound is equal to its objective, and the highest amongst these models. It also exhibits the best model fit in our experiments. Together, these results suggest that the predominant view of deterministic dropout as a good approximation to MC averaging is misleading. Rather, deterministic dropout is the best available approximation to the true objective.
****
Confidence Scores Make Instance-dependent Label-noise Learning Possible
Antonin Berthon,Bo Han,Gang Niu,Tongliang Liu,Masashi Sugiyama
Learning with noisy labels has drawn a lot of attention. In this area, most of recent works only consider class-conditional noise, where the label noise is independent of its input features. This noise model may not be faithful to many real-world applications. Instead, few pioneer works have studied instance-dependent noise, but these methods are limited to strong assumptions on noise models. To alleviate this issue, we introduce confidence-scored instance-dependent noise (CSIDN), where each instance-label pair is associated with a confidence score. The

confidence scores are sufficient to estimate the noise functions of each instanc e with minimal assumptions. Moreover, such scores can be easily and cheaply deri ved during the construction of the dataset through crowdsourcing or automatic an notation. To handle CSIDN, we design a benchmark algorithm termed instance-level forward correction. Empirical results on synthetic and real-world datasets demo nstrate the utility of our proposed method.
****

Gap-Aware Mitigation of Gradient Staleness
Saar Barkai,Ido Hakimi,Assaf Schuster
Cloud computing is becoming increasingly popular as a platform for distributed t raining of deep neural networks. Synchronous stochastic gradient descent (SSGD) suffers from substantial slowdowns due to stragglers if the environment is non-d edicated, as is common in cloud computing. Asynchronous SGD (ASGD) methods are i mmune to these slowdowns but are scarcely used due to gradient staleness, which encumbers the convergence process. Recent techniques have had limited success mi tigating the gradient staleness when scaling up to many workers (computing nodes). In this paper we define the Gap as a measure of gradient staleness and propo se Gap-Aware (GA), a novel asynchronous-distributed method that penalizes stale gradients linearly to the Gap and performs well even when scaling to large numbe rs of workers. Our evaluation on the CIFAR, ImageNet, and WikiText-103 datasets shows that GA outperforms the currently acceptable gradient penalization method, in final test accuracy. We also provide convergence rate proof for GA. Despite prior beliefs, we show that if GA is applied, momentum becomes beneficial in asy nchronous environments, even when the number of workers scales up.
****

Evaluating and Calibrating Uncertainty Prediction in Regression Tasks
Dan Levi,Liran Gispan,Niv Giladi,Ethan Fetaya
Predicting not only the target but also an accurate measure of uncertainty is im portant for many applications and in particular safety-critical ones. In this wo rk we study the calibration of uncertainty prediction for regression tasks which often arise in real-world systems. We show that the existing definition for cal ibration of a regression uncertainty [Kuleshov et al. 2018] has severe limitatio ns in distinguishing informative from non-informative uncertainty predictions. W e propose a new definition that escapes this caveat and an evaluation method usi ng a simple histogram-based approach inspired by reliability diagrams used in cl assification tasks. Our method clusters examples with similar uncertainty predic tion and compares the prediction with the empirical uncertainty on these example s. We also propose a simple scaling-based calibration that preforms well in our experimental tests. We show results on both a synthetic, controlled problem and on the object detection bounding-box regression task using the COCO and KITTI d atasets.
****

Ensemble Distribution Distillation
Andrey Malinin,Bruno Mlodozeniec,Mark Gales
Ensembles of models often yield improvements in system performance. These ensemb le approaches have also been empirically shown to yield robust measures of uncer tainty, and are capable of distinguishing between different forms of uncertainty . However, ensembles come at a computational and memory cost which may be prohib itive for many applications. There has been significant work done on the distill ation of an ensemble into a single model. Such approaches decrease computational cost and allow a single model to achieve an accuracy comparable to that of an e nsemble. However, information about the diversity of the ensemble, which can yie ld estimates of different forms of uncertainty, is lost. This work considers the novel task of Ensemble Distribution Distillation (EnD^2) - distilling the distr ibution of the predictions from an ensemble, rather than just the average predic tion, into a single model. EnD^2 enables a single model to retain both the impro ved classification performance of ensemble distillation as well as information a bout the diversity of the ensemble, which is useful for uncertainty estimation. A solution for EnD^2 based on Prior Networks, a class of models which allow a si ngle neural network to explicitly model a distribution over output distributions

, is proposed in this work. The properties of EnD^2 are investigated on both an artificial dataset, and on the CIFAR-10, CIFAR-100 and TinyImageNet datasets, where it is shown that EnD^2 can approach the classification performance of an ensemble, and outperforms both standard DNNs and Ensemble Distillation on the tasks of misclassification and out-of-distribution input detection.
****

Deformable Kernels: Adapting Effective Receptive Fields for Object Deformation
Hang Gao,Xizhou Zhu,Stephen Lin,Jifeng Dai
Convolutional networks are not aware of an object's geometric variations, which leads to inefficient utilization of model and data capacity. To overcome this issue, recent works on deformation modeling seek to spatially reconfigure the data towards a common arrangement such that semantic recognition suffers less from deformation. This is typically done by augmenting static operators with learned free-form sampling grids in the image space, dynamically tuned to the data and task for adapting the receptive field. Yet adapting the receptive field does not quite reach the actual goal -- what really matters to the network is the *effective* receptive field (ERF), which reflects how much each pixel contributes. It is thus natural to design other approaches to adapt the ERF directly during runtime. In this work, we instantiate one possible solution as Deformable Kernels (DKs), a family of novel and generic convolutional operators for handling object deformations by directly adapting the ERF while leaving the receptive field untouched. At the heart of our method is the ability to resample the original kernel space towards recovering the deformation of objects. This approach is justified with theoretical insights that the ERF is strictly determined by data sampling locations and kernel values. We implement DKs as generic drop-in replacements of rigid kernels and conduct a series of empirical studies whose results conform with our theories. Over several tasks and standard base models, our approach compares favorably against prior works that adapt during runtime. In addition, further experiments suggest a working mechanism orthogonal and complementary to previous works.
****

On the Tunability of Optimizers in Deep Learning
Prabhu Teja S*,Florian Mai*,Thijs Vogels,Martin Jaggi,Francois Fleuret
There is no consensus yet on the question whether adaptive gradient methods like Adam are easier to use than non-adaptive optimization methods like SGD. In this work, we fill in the important, yet ambiguous concept of ‘ease-of-use’ by defining an optimizer’s tunability: How easy is it to find good hyperparameter configurations using automatic random hyperparameter search? We propose a practical and universal quantitative measure for optimizer tunability that can form the basis for a fair optimizer benchmark. Evaluating a variety of optimizers on an extensive set of standard datasets and architectures, we find that Adam is the most tunable for the majority of problems, especially with a low budget for hyperparameter tuning.
****

Gradient Perturbation is Underrated for Differentially Private Convex Optimization
Da Yu,Huishuai Zhang,Wei Chen,Tie-yan Liu,Jian Yin
Gradient perturbation, widely used for differentially private optimization, injects noise at every iterative update to guarantee differential privacy. Previous work first determines the noise level that can satisfy the privacy requirement and then analyzes the utility of noisy gradient updates as in non-private case.
  In this paper, we explore how the privacy noise affects the optimization property. We show that for differentially private convex optimization, the utility guarantee of both DP-GD and DP-SGD is determined by an \emph{expected curvature} rather than the minimum curvature. The \emph{expected curvature} represents the average curvature over the optimization path, which is usually much larger than the minimum curvature and hence can help us achieve a significantly improved utility guarantee. By using the \emph{expected curvature}, our theory justifies the advantage of gradient perturbation over other perturbation methods and closes the gap between theory and practice. Extensive experiments on real world datasets

corroborate our theoretical findings.
****
VL-BERT: Pre-training of Generic Visual-Linguistic Representations
Weijie Su,Xizhou Zhu,Yue Cao,Bin Li,Lewei Lu,Furu Wei,Jifeng Dai
We introduce a new pre-trainable generic representation for visual-linguistic tasks, called Visual-Linguistic BERT (VL-BERT for short). VL-BERT adopts the simple yet powerful Transformer model as the backbone, and extends it to take both visual and linguistic embedded features as input. In it, each element of the input is either of a word from the input sentence, or a region-of-interest (RoI) from the input image. It is designed to fit for most of the visual-linguistic downstream tasks. To better exploit the generic representation, we pre-train VL-BERT on the massive-scale Conceptual Captions dataset, together with text-only corpus. Extensive empirical analysis demonstrates that the pre-training procedure can better align the visual-linguistic clues and benefit the downstream tasks, such as visual commonsense reasoning, visual question answering and referring expression comprehension. It is worth noting that VL-BERT achieved the first place of single model on the leaderboard of the VCR benchmark.
****
Credible Sample Elicitation by Deep Learning, for Deep Learning
Yang Liu,Zuyue Fu,Zhuoran Yang,Zhaoran Wang
It is important to collect credible training samples (x,y) for building data-intensive learning systems (e.g., a deep learning system). In the literature, there is a line of studies on eliciting distributional information from self-interested agents who hold a relevant information. Asking people to report complex distribution $p(x)$, though theoretically viable, is challenging in practice. This is primarily due to the heavy cognitive loads required for human agents to reason and report this high dimensional information. Consider the example where we are interested in building an image classifier via first collecting a certain category of high-dimensional image data. While classical elicitation results apply to eliciting a complex and generative (and continuous) distribution $p(x)$ for this image data, we are interested in eliciting samples $x_i \sim p(x)$ from agents. This paper introduces a deep learning aided method to incentivize credible sample contributions from selfish and rational agents. The challenge to do so is to design an incentive-compatible score function to score each reported sample to induce truthful reports, instead of an arbitrary or even adversarial one. We show that with accurate estimation of a certain f-divergence function we are able to achieve approximate incentive compatibility in eliciting truthful samples. We then present an efficient estimator with theoretical guarantee via studying the variational forms of f-divergence function. Our work complements the literature of information elicitation via introducing the problem of \emph{sample elicitation}. We also show a connection between this sample elicitation problem and f-GAN, and how this connection can help reconstruct an estimator of the distribution based on collected samples.
****
Neural Markov Logic Networks
Giuseppe Marra,Ond■ej Kuželka
We introduce Neural Markov Logic Networks (NMLNs), a statistical relational learning system that borrows ideas from Markov logic. Like Markov Logic Networks (MLNs), NMLNs are an exponential-family model for modelling distributions over possible worlds, but unlike MLNs, they do not rely on explicitly specified first-order logic rules. Instead, NMLNs learn an implicit representation of such rules as a neural network that acts as a potential function on fragments of the relational structure. Interestingly, any MLN can be represented as an NMLN. Similarly to recently proposed Neural theorem provers (NTPs) (Rocktaschel at al. 2017), NMLNs can exploit embeddings of constants but, unlike NTPs, NMLNs work well also in their absence. This is extremely important for predicting in settings other than the transductive one. We showcase the potential of NMLNs on knowledge-base completion tasks and on generation of molecular (graph) data.
****
Optimistic Exploration even with a Pessimistic Initialisation

Tabish Rashid,Bei Peng,Wendelin Boehmer,Shimon Whiteson
Optimistic initialisation is an effective strategy for efficient exploration in reinforcement learning (RL). In the tabular case, all provably efficient model-free algorithms rely on it. However, model-free deep RL algorithms do not use optimistic initialisation despite taking inspiration from these provably efficient tabular algorithms. In particular, in scenarios with only positive rewards, Q-values are initialised at their lowest possible values due to commonly used network initialisation schemes, a pessimistic initialisation. Merely initialising the network to output optimistic Q-values is not enough, since we cannot ensure that they remain optimistic for novel state-action pairs, which is crucial for exploration. We propose a simple count-based augmentation to pessimistically initialised Q-values that separates the source of optimism from the neural network. We show that this scheme is provably efficient in the tabular setting and extend it to the deep RL setting. Our algorithm, Optimistic Pessimistically Initialised Q-Learning (OPIQ), augments the Q-value estimates of a DQN-based agent with count-derived bonuses to ensure optimism during both action selection and bootstrapping. We show that OPIQ outperforms non-optimistic DQN variants that utilise a pseudocount-based intrinsic motivation in hard exploration tasks, and that it predicts optimistic estimates for novel state-action pairs.
****

Risk Averse Value Expansion for Sample Efficient and Robust Policy Learning
Bo Zhou,Fan Wang,Hongsheng Zeng,Hao Tian
Model-based Reinforcement Learning(RL) has shown great advantage in sample-efficiency, but suffers from poor asymptotic performance and high inference cost. A promising direction is to combine model-based reinforcement learning with model-free reinforcement learning, such as model-based value expansion(MVE). However, the previous methods do not take into account the stochastic character of the environment, thus still suffers from higher function approximation errors. As a result, they tend to fall behind the best model-free algorithms in some challenging scenarios. We propose a novel Hybrid-RL method, which is developed from MVE, namely the Risk Averse Value Expansion(RAVE). In the proposed method, we use an ensemble of probabilistic models for environment modeling to generate imaginative rollouts, based on which we further introduce the aversion of risks by seeking the lower confidence bound of the estimation. Experiments on different environments including MuJoCo and robo-school show that RAVE yields state-of-the-art performance. Also we found that it greatly prevented some catastrophic consequences such as falling down and thus reduced the variance of the rewards.
****

Certified Robustness for Top-k Predictions against Adversarial Perturbations via Randomized Smoothing
Jinyuan Jia,Xiaoyu Cao,Binghui Wang,Neil Zhenqiang Gong
It is well-known that classifiers are vulnerable to adversarial perturbations. To defend against adversarial perturbations, various certified robustness results have been derived. However, existing certified robustnesses are limited to top-1 predictions. In many real-world applications, top-k predictions are more relevant. In this work, we aim to derive certified robustness for top-k predictions. In particular, our certified robustness is based on randomized smoothing, which turns any classifier to a new classifier via adding noise to an input example. We adopt randomized smoothing because it is scalable to large-scale neural networks and applicable to any classifier. We derive a tight robustness in ℓ_2 norm for top-k predictions when using randomized smoothing with Gaussian noise. We find that generalizing the certified robustness from top-1 to top-k predictions faces significant technical challenges. We also empirically evaluate our method on CIFAR10 and ImageNet. For example, our method can obtain an ImageNet classifier with a certified top-5 accuracy of 62.8\% when the ℓ_2-norms of the adversarial perturbations are less than 0.5 (=127/255). Our code is publicly available at: \url{https://github.com/jjy1994/Certify_Topk}.
****

Generalized Domain Adaptation with Covariate and Label Shift CO-ALignment
Shuhan Tan,Xingchao Peng,Kate Saenko

Unsupervised knowledge transfer has a great potential to improve the generalizab
ility of deep models to novel domains. Yet the current literature assumes that t
he label distribution is domain-invariant and only aligns the covariate or vice
versa. In this paper, we explore the task of Generalized Domain Adaptation (GDA)
: How to transfer knowledge across different domains in the presence of both cov
ariate and label shift? We propose a covariate and label distribution CO-ALignme
nt (COAL) model to tackle this problem. Our model leverages prototype-based cond
itional alignment and label distribution estimation to diminish the covariate an
d label shifts, respectively. We demonstrate experimentally that when both types
 of shift exist in the data, COAL leads to state-of-the-art performance on sever
al cross-domain benchmarks.
****

LabelFool: A Trick in the Label Space
Yujia Liu,Tingting Jiang,Ming Jiang
It is widely known that well-designed perturbations can cause state-of-the-art m
achine learning classifiers to mis-label an image, with sufficiently small pertu
rbations that are imperceptible to the human eyes. However, by detecting the inc
onsistency between the image and wrong label, the human observer would be alerte
d of the attack. In this paper, we aim to design attacks that not only make clas
sifiers generate wrong labels, but also make the wrong labels imperceptible to h
uman observers. To achieve this, we propose an algorithm called LabelFool which
identifies a target label similar to the ground truth label and finds a perturba
tion of the image for this target label. We first find the target label for an i
nput image by a probability model, then move the input in the feature space towa
rds the target label. Subjective studies on ImageNet show that in the label spac
e, our attack is much less recognizable by human observers, while objective expe
rimental results on ImageNet show that we maintain similar performance in the im
age space as well as attack rates to state-of-the-art attack algorithms.
****

RGTI:Response generation via templates integration for End to End dialog
Yuxin Zhang,Songyan Liu
End-to-end models have achieved considerable success in task-oriented dialogue a
rea, but suffer from the challenges of (a) poor semantic control, and (b) little
 interaction with auxiliary information. In this paper, we propose a novel yet s
imple end-to-end model for response generation via mixed templates, which can ad
dress above challenges.
In our model, we retrieval candidate responses which contain abundant syntactic
and sequence information by dialogue semantic information related to dialogue hi
story. Then, we exploit candidate response attention to get templates which shou
ld be mentioned in response. Our model can integrate multi template information
to guide the decoder module how to generate response better. We show that our pr
oposed model learns useful templates information, which improves the performance
 of "how to say" and "what to say" in response generation. Experiments on the la
rge-scale Multiwoz dataset demonstrate the effectiveness of our proposed model,
which attain the state-of-the-art performance.
****

Towards Disentangling Non-Robust and Robust Components in Performance Metric
Yujun Shi,Benben Liao,Guangyong Chen,Yun Liu,Ming-ming Cheng,Jiashi Feng
The vulnerability to slight input perturbations is a worrying yet intriguing pro
perty of deep neural networks (DNNs). Though some efforts have been devoted to i
nvestigating the reason behind such adversarial behavior, the relation between s
tandard accuracy and adversarial behavior of DNNs is still little understood. In
 this work, we reveal such relation by first introducing a metric characterizing
 the standard performance of DNNs. Then we theoretically show this metric can be
 disentangled into an information-theoretic non-robust component that is related
 to adversarial behavior, and a robust component. Then, we show by experiments t
hat DNNs under standard training rely heavily on optimizing the non-robust compo
nent in achieving decent performance. We also demonstrate current state-of-the-a
rt adversarial training algorithms indeed try to robustify DNNs by preventing th
em from using the non-robust component to distinguish samples from different cat

egories. Based on our findings, we take a step forward and point out the possibl e direction of simultaneously achieving decent standard generalization and adver sarial robustness. It is hoped that our theory can further inspire the community to make more interesting discoveries about the relation between standard accura cy and adversarial robustness of DNNs.
****

A Mechanism of Implicit Regularization in Deep Learning
Masayoshi Kubo,Genki Sugiura,Kenta Shinzato,Momose Oyama
Despite a lot of theoretical efforts, very little is known about mechanisms of i mplicit regularization by which the low complexity contributes to generalization in deep learning. In particular, causality between the generalization performan ce, implicit regularization and nonlinearity of activation functions is one of t he basic mysteries of deep neural networks (DNNs). In this work, we introduce a novel technique for DNNs called random walk analysis and reveal a mechanism of t he implicit regularization caused by nonlinearity of ReLU activation. Surprising ly, our theoretical results suggest that the learned DNNs interpolate almost lin early between data points, which leads to the low complexity solutions in the ov er-parameterized regime. As a result, we prove that stochastic gradient descent can learn a class of continuously differentiable functions with generalization b ounds of the order of $O(n^{-2})$ (n: the number of samples). Furthermore, our analysis is independent of the kernel methods, including neural tangent kernels .
****

Feature-map-level Online Adversarial Knowledge Distillation
Inseop Chung,SeongUk Park,Jangho Kim,Nojun Kwak
Feature maps contain rich information about image intensity and spatial correlat ion. However, previous online knowledge distillation methods only utilize the cl ass probabilities. Thus in this paper, we propose an online knowledge distillati on method that transfers not only the knowledge of the class probabilities but a lso that of the feature map using the adversarial training framework. We train m ultiple networks simultaneously by employing discriminators to distinguish the f eature map distributions of different networks. Each network has its correspondi ng discriminator which discriminates the feature map from its own as fake while classifying that of the other network as real. By training a network to fool the corresponding discriminator, it can learn the other network’s feature map distr ibution. Discriminators and networks are trained concurrently in a minimax two-p layer game. Also, we propose a novel cyclic learning scheme for training more th an two networks together. We have applied our method to various network architec tures on the classification task and discovered a significant improvement of per formance especially in the case of training a pair of a small network and a larg e one.
****

Optimising Neural Network Architectures for Provable Adversarial Robustness
Henry Gouk,Timothy M. Hospedales
Existing Lipschitz-based provable defences to adversarial examples only cover th e L2 threat model. We introduce the first bound that makes use of Lipschitz cont inuity to provide a more general guarantee for threat models based on any p-norm . Additionally, a new strategy is proposed for designing network architectures t hat exhibit superior provable adversarial robustness over conventional convoluti onal neural networks. Experiments are conducted to validate our theoretical cont ributions, show that the assumptions made during the design of our novel archite cture hold in practice, and quantify the empirical robustness of several Lipschi tz-based adversarial defence methods.
****

Recurrent Independent Mechanisms
Anirudh Goyal,Alex Lamb,Shagun Sodhani,Jordan Hoffmann,Sergey Levine,Yoshua Beng io,Bernhard Scholkopf
Learning modular structures which reflect the dynamics of the environment can le ad to better generalization and robustness to changes which only affect a few of the underlying causes. We propose Recurrent Independent Mechanisms (RIMs), a ne

w recurrent architecture in which multiple groups of recurrent cells operate wit
h nearly independent transition dynamics, communicate only sparingly through the
 bottleneck of attention, and are only updated at time steps where they are most
 relevant. We show that this leads to specialization amongst the RIMs, which in
 turn allows for dramatically improved generalization on tasks where some factor
s of variation differ systematically between training and evaluation.
****

An Explicitly Relational Neural Network Architecture
Murray Shanahan,Kyriacos Nikiforou,Antonia Creswell,Christos Kaplanis,David Barr
ett,Marta Garnelo
With a view to bridging the gap between deep learning and symbolic AI, we presen
t a novel end-to-end neural network architecture that learns to form proposition
al representations with an explicitly relational structure from raw pixel data.
In order to evaluate and analyse the architecture, we introduce a family of simp
le visual relational reasoning tasks of varying complexity. We show that the pro
posed architecture, when pre-trained on a curriculum of such tasks, learns to ge
nerate reusable representations that better facilitate subsequent learning on pr
eviously unseen tasks when compared to a number of baseline architectures. The w
orkings of a successfully trained model are visualised to shed some light on how
 the architecture functions.
****

Branched Multi-Task Networks: Deciding What Layers To Share
Simon Vandenhende,Stamatios Georgoulis,Bert De Brabandere,Luc Van Gool
In the context of multi-task learning, neural networks with branched architectur
es have often been employed to jointly tackle the tasks at hand. Such ramified n
etworks typically start with a number of shared layers, after which different ta
sks branch out into their own sequence of layers. Understandably, as the number
of possible network configurations is combinatorially large, deciding what layer
s to share and where to branch out becomes cumbersome. Prior works have either r
elied on ad hoc methods to determine the level of layer sharing, which is subopt
imal, or utilized neural architecture search techniques to establish the network
 design, which is considerably expensive. In this paper, we go beyond these limi
tations and propose a principled approach to automatically construct branched mu
lti-task networks, by leveraging the employed tasks' affinities. Given a specifi
c budget, i.e. number of learnable parameters, the proposed approach generates a
rchitectures, in which shallow layers are task-agnostic, whereas deeper ones gra
dually grow more task-specific. Extensive experimental analysis across numerous,
 diverse multi-tasking datasets shows that, for a given budget, our method consi
stently yields networks with the highest performance, while for a certain perfor
mance threshold it requires the least amount of learnable parameters.
****

MxPool: Multiplex Pooling for Hierarchical Graph Representation Learning
Yanyan Liang,Yanfeng Zhang,Fangjing Wang,Qian Xu
Graphs are known to have complicated structures and have myriad applications. Ho
w to utilize deep learning methods for graph classification tasks has attracted
considerable research attention in the past few years. Two properties of graph d
ata have imposed significant challenges on existing graph learning techniques. (
1) Diversity: each graph has a variable size of unordered nodes and diverse node
/edge types. (2) Complexity: graphs have not only node/edge features but also co
mplex topological features. These two properties motivate us to use multiplex st
ructure to learn graph features in a diverse way. In this paper, we propose a si
mple but effective approach, MxPool, which concurrently uses multiple graph conv
olution networks and graph pooling networks to build hierarchical learning struc
ture for graph representation learning tasks. Our experiments on numerous graph
classification benchmarks show that our MxPool has marked superiority over other
 state-of-the-art graph representation learning methods. For example, MxPool ach
ieves 92.1% accuracy on the D&D dataset while the second best method DiffPool on
ly achieves 80.64% accuracy.
****

Mixture-of-Experts Variational Autoencoder for clustering and generating from si

milarity-based representations
Andreas Kopf,Vincent Fortuin,Vignesh Ram Somnath,Manfred Claassen
Clustering high-dimensional data, such as images or biological measurements, is a long-standing problem and has been studied extensively. Recently, Deep Clustering gained popularity due to the non-linearity of neural networks, which allows for flexibility in fitting the specific peculiarities of complex data. Here we introduce the Mixture-of-Experts Similarity Variational Autoencoder (MoE-Sim-VAE), a novel generative clustering model. The model can learn multi-modal distributions of high-dimensional data and use these to generate realistic data with high efficacy and efficiency. MoE-Sim-VAE is based on a Variational Autoencoder (VAE), where the decoder consists of a Mixture-of-Experts (MoE) architecture. This specific architecture allows for various modes of the data to be automatically learned by means of the experts. Additionally, we encourage the latent representation of our model to follow a Gaussian mixture distribution and to accurately represent the similarities between the data points. We assess the performance of our model on synthetic data, the MNIST benchmark data set, and a challenging real-world task of defining cell subpopulations from mass cytometry (CyTOF) measurements on hundreds of different datasets. MoE-Sim-VAE exhibits superior clustering performance on all these tasks in comparison to the baselines and we show that the MoE architecture in the decoder reduces the computational cost of sampling specific data modes with high fidelity.
****
Temporal Difference Weighted Ensemble For Reinforcement Learning
Takuma Seno,Michita Imai
Combining multiple function approximators in machine learning models typically leads to better performance and robustness compared with a single function. In reinforcement learning, ensemble algorithms such as an averaging method and a majority voting method are not always optimal, because each function can learn fundamentally different optimal trajectories from exploration. In this paper, we propose a Temporal Difference Weighted (TDW) algorithm, an ensemble method that adjusts weights of each contribution based on accumulated temporal difference errors. The advantage of this algorithm is that it improves ensemble performance by reducing weights of Q-functions unfamiliar with current trajectories. We provide experimental results for Gridworld tasks and Atari tasks that show significant performance improvements compared with baseline algorithms.
****
Effect of top-down connections in Hierarchical Sparse Coding
Victor Boutin,Angelo Franciosini,Franck Ruffier,Laurent Perrinet
Hierarchical Sparse Coding (HSC) is a powerful model to efficiently represent multi-dimensional, structured data such as images. The simplest solution to solve this computationally hard problem is to decompose it into independent layerwise subproblems. However, neuroscientific evidence would suggest inter-connecting these subproblems as in the Predictive Coding (PC) theory, which adds top-down connections between consecutive layers. In this study, a new model called Sparse Deep Predictive Coding (SDPC) is introduced to assess the impact of this inter-layer feedback connection. In particular, the SDPC is compared with a Hierarchical Lasso (Hi-La) network made out of a sequence of Lasso layers. A 2-layered SDPC and a Hi-La networks are trained on 3 different databases and with different sparsity parameters on each layer. First, we show that the overall prediction error generated by SDPC is lower thanks to the feedback mechanism as it transfers prediction error between layers. Second, we demonstrate that the inference stage of the SDPC is faster to converge than for the Hi-La model. Third, we show that the SDPC also accelerates the learning process. Finally, the qualitative analysis of both models dictionaries, supported by their activation probability, show that the SDPC features are more generic and informative.
****
Compressive Recovery Defense: A Defense Framework for ℓ_0, ℓ_2 and ℓ_∞ norm attacks.
Jasjeet Dhaliwal,Kyle Hambrook
We provide recovery guarantees for compressible signals that have been corrupted

with noise and extend the framework introduced in \cite{bafna2018thwarting} to defend neural networks against ℓ_0, ℓ_2, and ℓ_{∞}-norm attacks. In the case of ℓ_0-norm noise, we provide recovery guarantees for Iterative Hard Thresholding (IHT) and Basis Pursuit (BP). For ℓ_2-norm bounded noise, we provide recovery guarantees for BP, and for the case of ℓ_∞-norm bounded noise, we provide recovery guarantees for Dantzig Selector (DS). These guarantees theoretically bolster the defense framework introduced in \cite{bafna2018thwarting} for defending neural networks against adversarial inputs. Finally, we experimentally demonstrate the effectiveness of this defense framework against an array of ℓ_0, ℓ_2 and ℓ_∞-norm attacks.
****

Match prediction from group comparison data using neural networks
Sunghyun Kim,Minje jang,Changho Suh
We explore the match prediction problem where one seeks to estimate the likelihood of a group of M items preferred over another, based on partial group comparison data. Challenges arise in practice. As existing state-of-the-art algorithms are tailored to certain statistical models, we have different best algorithms across distinct scenarios. Worse yet, we have no prior knowledge on the underlying model for a given scenario. These call for a unified approach that can be universally applied to a wide range of scenarios and achieve consistently high performances. To this end, we incorporate deep learning architectures so as to reflect the key structural features that most state-of-the-art algorithms, some of which are optimal in certain settings, share in common. This enables us to infer hidden models underlying a given dataset, which govern in-group interactions and statistical patterns of comparisons, and hence to devise the best algorithm tailored to the dataset at hand. Through extensive experiments on synthetic and real-world datasets, we evaluate our framework in comparison to state-of-the-art algorithms. It turns out that our framework consistently leads to the best performance across all datasets in terms of cross entropy loss and prediction accuracy, while the state-of-the-art algorithms suffer from inconsistent performances across different datasets. Furthermore, we show that it can be easily extended to attain satisfactory performances in rank aggregation tasks, suggesting that it can be adaptable for other tasks as well.
****

Identifying through Flows for Recovering Latent Representations
Shen Li,Bryan Hooi,Gim Hee Lee
Identifiability, or recovery of the true latent representations from which the observed data originates, is de facto a fundamental goal of representation learning. Yet, most deep generative models do not address the question of identifiability, and thus fail to deliver on the promise of the recovery of the true latent sources that generate the observations. Recent work proposed identifiable generative modelling using variational autoencoders (iVAE) with a theory of identifiability. Due to the intractablity of KL divergence between variational approximate posterior and the true posterior, however, iVAE has to maximize the evidence lower bound (ELBO) of the marginal likelihood, leading to suboptimal solutions in both theory and practice. In contrast, we propose an identifiable framework for estimating latent representations using a flow-based model (iFlow). Our approach directly maximizes the marginal likelihood, allowing for theoretical guarantees on identifiability, thereby dispensing with variational approximations. We derive its optimization objective in analytical form, making it possible to train iFlow in an end-to-end manner. Simulations on synthetic data validate the correctness and effectiveness of our proposed method and demonstrate its practical advantages over other existing methods.
****

Robust training with ensemble consensus
Jisoo Lee,Sae-Young Chung
Since deep neural networks are over-parameterized, they can memorize noisy examples. We address such a memorization issue in the presence of label noise. From the fact that deep neural networks cannot generalize to neighborhoods of memorized features, we hypothesize that noisy examples do not consistently incur small l

osses on the network under a certain perturbation. Based on this, we propose a n ovel training method called Learning with Ensemble Consensus (LEC) that prevents overfitting to noisy examples by removing them based on the consensus of an ens emble of perturbed networks. One of the proposed LECs, LTEC outperforms the curr ent state-of-the-art methods on noisy MNIST, CIFAR-10, and CIFAR-100 in an effic ient manner.

****

BRIDGING ADVERSARIAL SAMPLES AND ADVERSARIAL NETWORKS

Faqiang Liu,Mingkun Xu,Guoqi Li,Jing Pei,Luping Shi

Generative adversarial networks have achieved remarkable performance on various tasks but suffer from sensitivity to hyper-parameters, training instability, and mode collapse. We find that this is partly due to gradient given by non-robust discriminator containing non-informative adversarial noise, which can hinder gen erator from catching the pattern of real samples. Inspired by defense against ad versarial samples, we introduce adversarial training of discriminator on real sa mples that does not exist in classic GANs framework to make adversarial training symmetric, which can balance min-max game and make discriminator more robust. R obust discriminator can give more informative gradient with less adversarial noi se, which can stabilize training and accelerate convergence. We validate the pro posed method on image generation tasks with varied network architectures quantit atively. Experiments show that training stability, perceptual quality, and diver sity of generated samples are consistently improved with small additional traini ng computation cost.

****

Unsupervised-Learning of time-varying features

Henrik Høeg,Matthias Brix,Oswin Krause

We present an architecture based on the conditional Variational Autoencoder to l earn a representation
of transformations in time-sequence data. The model is constructed in a way that allows to identify sub-spaces of features indicating changes between frames wit hout learning features that are constant within a time-sequence. Therefore, the approach disentangles content from transformations. Different model-architecture s are applied to affine image-transformations on MNIST as well as a car-racing v ideo-game task.
Results show that the model discovers relevant parameterizations, however, model architecture has a major impact on the feature-space. It turns out, that there is an advantage of only learning features describing change of state between ima ges, over learning the states of the images at each frame. In this case, we do n ot only achieve higher accuracy but also more interpretable linear features. Our results also uncover the need for model architectures that combine global trans formations with convolutional architectures.

****

Self-Adversarial Learning with Comparative Discrimination for Text Generation

Wangchunshu Zhou,Tao Ge,Ke Xu,Furu Wei,Ming Zhou

Conventional Generative Adversarial Networks (GANs) for text generation tend to have issues of reward sparsity and mode collapse that affect the quality and div ersity of generated samples. To address the issues, we propose a novel self-adve rsarial learning (SAL) paradigm for improving GANs' performance in text generati on. In contrast to standard GANs that use a binary classifier as its discriminat or to predict whether a sample is real or generated, SAL employs a comparative d iscriminator which is a pairwise classifier for comparing the text quality betwe en a pair of samples. During training, SAL rewards the generator when its curren tly generated sentence is found to be better than its previously generated sampl es. This self-improvement reward mechanism allows the model to receive credits m ore easily and avoid collapsing towards the limited number of real samples, whic h not only helps alleviate the reward sparsity issue but also reduces the risk o f mode collapse. Experiments on text generation benchmark datasets show that our proposed approach substantially improves both the quality and the diversity, an d yields more stable performance compared to the previous GANs for text generati on.

****
A General Upper Bound for Unsupervised Domain Adaptation
Dexuan Zhang,Tatsuya Harada
In this work, we present a novel upper bound of target error to address the prob lem for unsupervised domain adaptation. Recent studies reveal that a deep neural network can learn transferable features which generalize well to novel tasks. F urthermore, Ben-David et al. (2010) provide an upper bound for target error whe n transferring the knowledge, which can be summarized as minimizing the source e rror and distance between marginal distributions simultaneously. However, commo n methods based on the theory usually ignore the joint error such that samples f rom different classes might be mixed together when matching marginal distributio n. And in such case, no matter how we minimize the marginal discrepancy, the tar get error is not bounded due to an increasing joint error. To address this probl em, we propose a general upper bound taking joint error into account, such that the undesirable case can be properly penalized. In addition, we utilize constrai ned hypothesis space to further formalize a tighter bound as well as a novel cro ss margin discrepancy to measure the dissimilarity between hypotheses which alle viates instability during adversarial learning. Extensive empirical evidence sho ws that our proposal outperforms related approaches in image classification erro r rates on standard domain adaptation benchmarks.
****
Vid2Game: Controllable Characters Extracted from Real-World Videos
Oran Gafni,Lior Wolf,Yaniv Taigman
We extract a controllable model from a video of a person performing a certain ac tivity. The model generates novel image sequences of that person, according to u ser-defined control signals, typically marking the displacement of the moving bo dy. The generated video can have an arbitrary background, and effectively captur e both the dynamics and appearance of the person.

The method is based on two networks. The first maps a current pose, and a singl e-instance control signal to the next pose. The second maps the current pose, th e new pose, and a given background, to an output frame. Both networks include mu ltiple novelties that enable high-quality performance. This is demonstrated on m ultiple characters extracted from various videos of dancers and athletes.
****
Action Semantics Network: Considering the Effects of Actions in Multiagent Syste ms
Weixun Wang,Tianpei Yang,Yong Liu,Jianye Hao,Xiaotian Hao,Yujing Hu,Yingfeng Che n,Changjie Fan,Yang Gao
In multiagent systems (MASs), each agent makes individual decisions but all of t hem contribute globally to the system evolution. Learning in MASs is difficult s ince each agent's selection of actions must take place in the presence of other co-learning agents. Moreover, the environmental stochasticity and uncertainties increase exponentially with the increase in the number of agents. Previous works borrow various multiagent coordination mechanisms into deep learning architectu re to facilitate multiagent coordination. However, none of them explicitly consi der action semantics between agents that different actions have different influe nces on other agents. In this paper, we propose a novel network architecture, na med Action Semantics Network (ASN), that explicitly represents such action seman tics between agents. ASN characterizes different actions' influence on other age nts using neural networks based on the action semantics between them. ASN can be easily combined with existing deep reinforcement learning (DRL) algorithms to b oost their performance. Experimental results on StarCraft II micromanagement and Neural MMO show ASN significantly improves the performance of state-of-the-art DRL approaches compared with several network architectures.
****
Growing Action Spaces
Gregory Farquhar,Laura Gustafson,Zeming Lin,Shimon Whiteson,Nicolas Usunier,Gabr iel Synnaeve
In complex tasks, such as those with large combinatorial action spaces, random e

xploration may be too inefficient to achieve meaningful learning progress. In th is work, we use a curriculum of progressively growing action spaces to accelerat e learning. We assume the environment is out of our control, but that the agent may set an internal curriculum by initially restricting its action space. Our ap proach uses off-policy reinforcement learning to estimate optimal value function s for multiple action spaces simultaneously and efficiently transfers data, val ue estimates, and state representations from restricted action spaces to the ful l task. We show the efficacy of our approach in proof-of-concept control tasks a nd on challenging large-scale StarCraft micromanagement tasks with large, multi- agent action spaces.
****

Learning Generative Image Object Manipulations from Language Instructions
Martin Längkvist,Andreas Persson,Amy Loutfi
The use of adequate feature representations is essential for achieving high perf ormance in high-level human cognitive tasks in computational modeling. Recent de velopments in deep convolutional and recurrent neural networks architectures ena ble learning powerful feature representations from both images and natural langu age text. Besides, other types of networks such as Relational Networks (RN) can learn relations between objects and Generative Adversarial Networks (GAN) have s hown to generate realistic images. In this paper, we combine these four techniqu es to acquire a shared feature representation of the relation between objects in an input image and an object manipulation action description in the form of hum an language encodings to generate an image that shows the resulting end-effect t he action would have on a computer-generated scene. The system is trained and ev aluated on a simulated dataset and experimentally used on real-world photos.
****

Discourse-Based Evaluation of Language Understanding
Damien Sileo,Tim Van-De-Cruys,Camille Pradel,Philippe Muller
New models for natural language understanding have made unusual progress recentl y, leading to claims of universal text representations. However, current benchma rks are predominantly targeting semantic phenomena; we make the case that disco urse and pragmatics need to take center stage in the evaluation of natural langu age understanding.
We introduce DiscEval, a new benchmark for the evaluation of natural language un derstanding, that unites 11 discourse-focused evaluation datasets.
DiscEval can be used as supplementary training data in a multi-task learning set up, and is publicly available, alongside the code for gathering and preprocessin g the datasets.
Using our evaluation suite, we show that natural language inference, a widely us ed pretraining task, does not result in genuinely universal representations, whi ch opens a new challenge for multi-task learning.
****

Learning Efficient Parameter Server Synchronization Policies for Distributed SGD
Rong Zhu,Sheng Yang,Andreas Pfadler,Zhengping Qian,Jingren Zhou
We apply a reinforcement learning (RL) based approach to learning optimal synchr onization policies used for Parameter Server-based distributed training of machi ne learning models with Stochastic Gradient Descent (SGD). Utilizing a formal sy nchronization policy description in the PS-setting, we are able to derive a suit able and compact description of states and actions, allowing us to efficiently u se the standard off-the-shelf deep Q-learning algorithm. As a result, we are abl e to learn synchronization policies which generalize to different cluster enviro nments, different training datasets and small model variations and (most importa ntly) lead to considerable decreases in training time when compared to standard policies such as bulk synchronous parallel (BSP), asynchronous parallel (ASP), o r stale synchronous parallel (SSP). To support our claims we present extensive n umerical results obtained from experiments performed in simulated cluster enviro nments. In our experiments training time is reduced by 44 on average and learned policies generalize to multiple unseen circumstances.
****

Relational State-Space Model for Stochastic Multi-Object Systems

Fan Yang,Ling Chen,Fan Zhou,Yusong Gao,Wei Cao

Real-world dynamical systems often consist of multiple stochastic subsystems that interact with each other. Modeling and forecasting the behavior of such dynamics are generally not easy, due to the inherent hardness in understanding the complicated interactions and evolutions of their constituents. This paper introduces the relational state-space model (R-SSM), a sequential hierarchical latent variable model that makes use of graph neural networks (GNNs) to simulate the joint state transitions of multiple correlated objects. By letting GNNs cooperate with SSM, R-SSM provides a flexible way to incorporate relational information into the modeling of multi-object dynamics. We further suggest augmenting the model with normalizing flows instantiated for vertex-indexed random variables and propose two auxiliary contrastive objectives to facilitate the learning. The utility of R-SSM is empirically evaluated on synthetic and real time series datasets.

****

TSInsight: A local-global attribution framework for interpretability in time-series data

Shoaib Ahmed Siddiqui,Dominique Mercier,Andreas Dengel,Sheraz Ahmed

With the rise in employment of deep learning methods in safety-critical scenarios, interpretability is more essential than ever before. Although many different directions regarding interpretability have been explored for visual modalities, time-series data has been neglected with only a handful of methods tested due to their poor intelligibility. We approach the problem of interpretability in a novel way by proposing TSInsight where we attach an auto-encoder with a sparsity-inducing norm on its output to the classifier and fine-tune it based on the gradients from the classifier and a reconstruction penalty. The auto-encoder learns to preserve features that are important for the prediction by the classifier and suppresses the ones that are irrelevant i.e. serves as a feature attribution method to boost interpretability. In other words, we ask the network to only reconstruct parts which are useful for the classifier i.e. are correlated or causal for the prediction. In contrast to most other attribution frameworks, TSInsight is capable of generating both instance-based and model-based explanations. We evaluated TSInsight along with other commonly used attribution methods on a range of different time-series datasets to validate its efficacy. Furthermore, we analyzed the set of properties that TSInsight achieves out of the box including adversarial robustness and output space contraction. The obtained results advocate that TSInsight can be an effective tool for the interpretability of deep time-series models.

****

OPTIMAL TRANSPORT, CYCLEGAN, AND PENALIZED LS FOR UNSUPERVISED LEARNING IN INVERSE PROBLEMS

Byeongsu Sim,Gyutaek Oh,Sungjun Lim,and Jong Chul Ye

The penalized least squares (PLS) is a classic approach to inverse problems, where a regularization term is added to stabilize the solution. Optimal transport (OT) is another mathematical framework for computer vision tasks by providing means to transport one measure to another at minimal cost. Cycle-consistent generative adversarial network (cycleGAN) is a recent extension of GAN to learn target distributions with less mode collapsing behavior. Although similar in that no supervised training is required, the algorithms look different, so the mathematical relationship between these approaches is not clear. In this article, we provide an important advance to unveil the missing link. Specifically, we reveal that a cycleGAN architecture can be derived as a dual formulation of the optimal transport problem, if the PLS with a deep learning penalty is used as a transport cost between the two probability measures from measurements and unknown images. This suggests that cycleGAN can be considered as stochastic generalization of classical PLS approaches.

Our derivation is so general that various types of cycleGAN architecture can be easily derived by merely changing the transport cost. As proofs of concept, this paper provides novel cycleGAN architecture for unsupervised learning in accelerated MRI and deconvolution microscopy problems, which confirm the efficacy and the flexibility of the theory.

****
Structural Language Models for Any-Code Generation
Uri Alon,Roy Sadaka,Omer Levy,Eran Yahav
We address the problem of Any-Code Generation (AnyGen) - generating code without any restriction on the vocabulary or structure. The state-of-the-art in this problem is the sequence-to-sequence (seq2seq) approach, which treats code as a sequence and does not leverage any structural information. We introduce a new approach to AnyGen that leverages the strict syntax of programming languages to model a code snippet as tree structural language modeling (SLM). SLM estimates the probability of the program's abstract syntax tree (AST) by decomposing it into a product of conditional probabilities over its nodes. We present a neural model that computes these conditional probabilities by considering all AST paths leading to a target node. Unlike previous structural techniques that have severely restricted the kinds of expressions that can be generated, our approach can generate arbitrary expressions in any programming language. Our model significantly outperforms both seq2seq and a variety of existing structured approaches in generating Java and C# code. We make our code, datasets, and models available online.
****
Simple and Effective Stochastic Neural Networks
Tianyuan Yu,Yongxin Yang,Da Li,Timothy Hospedales,Tao Xiang
Stochastic neural networks (SNNs) are currently topical, with several paradigms being actively investigated including dropout, Bayesian neural networks, variational information bottleneck (VIB) and noise regularized learning. These neural network variants impact several major considerations, including generalization, network compression, and robustness against adversarial attack and label noise. However, many existing networks are complicated and expensive to train, and/or only address one or two of these practical considerations. In this paper we propose a simple and effective stochastic neural network (SE-SNN) architecture for discriminative learning by directly modeling activation uncertainty and encouraging high activation variability. Compared to existing SNNs, our SE-SNN is simpler to implement and faster to train, and produces state of the art results on network compression by pruning, adversarial defense and learning with label noise.
****
Robust Reinforcement Learning with Wasserstein Constraint
Linfang Hou,Liang Pang,Xin Hong,Yanyan Lan,Zhiming Ma,Dawei Yin
Robust Reinforcement Learning aims to find the optimal policy with some degree of robustness to environmental dynamics. Existing learning algorithms usually enable the robustness though disturbing the current state or simulated environmental parameters in a heuristic way, which lack quantified robustness to the system dynamics (i.e. transition probability). To overcome this issue, we leverage Wasserstein distance to measure the disturbance to the reference transition probability. With Wasserstein distance, we are able to connect transition probability disturbance to the state disturbance, and reduces an infinite-dimensional optimization problem to a finite-dimensional risk-aware problem. Through the derived risk-aware optimal Bellman equation, we first show the existence of optimal robust policies, provide a sensitivity analysis for the perturbations, and then design a novel robust learning algorithm—WassersteinRobustAdvantageActor-Critic algorithm (WRA2C). The effectiveness of the proposed algorithm is verified in theCart-Pole environment.
****
Cross-Iteration Batch Normalization
Zhuliang Yao,Yue Cao,Shuxin Zheng,Gao Huang,Stephen Lin,Jifeng Dai
A well-known issue of Batch Normalization is its significantly reduced effectiveness in the case of small mini-batch sizes. When a mini-batch contains few examples, the statistics upon which the normalization is defined cannot be reliably estimated from it during a training iteration. To address this problem, we present Cross-Iteration Batch Normalization (CBN), in which examples from multiple recent iterations are jointly utilized to enhance estimation quality. A challenge of computing statistics over multiple iterations is that the network activations from different iterations are not comparable to each other due to changes in net

work weights. We thus compensate for the network weight changes via a proposed t echnique based on Taylor polynomials, so that the statistics can be accurately e stimated and batch normalization can be effectively applied. On object detection and image classification with small mini-batch sizes, CBN is found to outperfor m the original batch normalization and a direct calculation of statistics over p revious iterations without the proposed compensation technique.
****
Model Ensemble-Based Intrinsic Reward for Sparse Reward Reinforcement Learning
Giseung Park,Whiyoung Jung,Sungho Choi,Youngchul Sung
In this paper, a new intrinsic reward generation method for sparse-reward reinfo rcement learning is proposed based on an ensemble of dynamics models. In the pro posed method, the mixture of multiple dynamics models is used to approximate the true unknown transition probability, and the intrinsic reward is designed as th e minimum of the surprise seen from each dynamics model to the mixture of the dy namics models. In order to show the effectiveness of the proposed intrinsic rewa rd generation method, a working algorithm is constructed by combining the propos ed intrinsic reward generation method with the proximal policy optimization (PPO) algorithm. Numerical results show that for representative locomotion tasks, th e proposed model-ensemble-based intrinsic reward generation method outperforms t he previous methods based on a single dynamics model.
****
The Effect of Residual Architecture on the Per-Layer Gradient of Deep Networks
Etai Littwin,Lior Wolf
A critical part of the training process of neural networks takes place in the ve ry first gradient steps post initialization. In this work, we study the connecti on between the network's architecture and initialization parameters, to the stat istical properties of the gradient in random fully connected ReLU networks, thro ugh the study of the the Jacobian. We compare three types of architectures: vani lla networks, ResNets and DenseNets. The later two, as we show, preserve the var iance of the gradient norm through arbitrary depths when initialized properly, which prevents exploding or decaying gradients at deeper layers. In addition, we show that the statistics of the per layer gradient norm is a function of the ar chitecture and the layer's size, but surprisingly not the layer's depth.

This depth invariant result is surprising in light of the literature results tha t state that the norm of the layer's activations grows exponentially with the sp ecific layer's depth. Experimental support is given in order to validate our the oretical results and to reintroduce concatenated ReLU blocks, which, as we show, present better initialization properties than ReLU blocks in the case of fully connected networks.
****
Prune or quantize? Strategy for Pareto-optimally low-cost and accurate CNN
Kengo Nakata,Daisuke Miyashita,Asuka Maki,Fumihiko Tachibana,Shinichi Sasaki,Jun Deguchi
Pruning and quantization are typical approaches to reduce the computational cost of CNN inference. Although the idea to combine them together seems natural, it is being unexpectedly difficult to figure out the resultant effect of the combin ation unless measuring the performance on a certain hardware which a user is goi ng to use. This is because the benefits of pruning and quantization strongly dep end on the hardware architecture where the model is executed. For example, a CPU -like architecture without any parallelization may fully exploit the reduction o f computations by unstructured pruning for speeding up, but a GPU-like massive p arallel architecture would not. Besides, there have been emerging proposals of n ovel hardware architectures such as one supporting variable bit precision quanti zation. From an engineering viewpoint, optimization for each hardware architectu re is useful and important in practice, but this is quite a brute-force approach . Therefore, in this paper, we first propose hardware-agnostic metric to measure the computational cost. And using the metric, we demonstrate that Pareto-optima l performance, where the best accuracy is obtained at a given computational cost , is achieved when a slim model with smaller number of parameters is quantized m

oderately rather than a fat model with huge number of parameters is quantized to extremely low bit precision such as binary or ternary. Furthermore, we empirically found the possible quantitative relation between the proposed metric and the signal to noise ratio during SGD training, by which the information obtained during SGD training provides the optimal policy of quantization and pruning. We show the Pareto frontier is improved by 4 times in post-training quantization scenario based on these findings. These findings are available not only to improve the Pareto frontier for accuracy vs. computational cost, but also give us some new insights on deep neural network.
****
Graph Residual Flow for Molecular Graph Generation
Shion Honda,Hirotaka Akita,Katsuhiko Ishiguro,Toshiki Nakanishi,Kenta Oono
Statistical generative models for molecular graphs attract attention from many researchers from the fields of bio- and chemo-informatics. Among these models, invertible flow-based approaches are not fully explored yet. In this paper, we propose a powerful invertible flow for molecular graphs, called Graph Residual Flow (GRF). The GRF is based on residual flows, which are known for more flexible and complex non-linear mappings than traditional coupling flows. We theoretically derive non-trivial conditions such that GRF is invertible, and present a way of keeping the entire flows invertible throughout the training and sampling. Experimental results show that a generative model based on the proposed GRF achieve comparable generation performance, with much smaller number of trainable parameters compared to the existing flow-based model.
****
Piecewise linear activations substantially shape the loss surfaces of neural networks
Fengxiang He,Bohan Wang,Dacheng Tao
Understanding the loss surface of a neural network is fundamentally important to the understanding of deep learning. This paper presents how piecewise linear activation functions substantially shape the loss surfaces of neural networks. We first prove that {\it the loss surfaces of many neural networks have infinite spurious local minima} which are defined as the local minima with higher empirical risks than the global minima. Our result demonstrates that the networks with piecewise linear activations possess substantial differences to the well-studied linear neural networks. This result holds for any neural network with arbitrary depth and arbitrary piecewise linear activation functions (excluding linear functions) under most loss functions in practice. Essentially, the underlying assumptions are consistent with most practical circumstances where the output layer is narrower than any hidden layer. In addition, the loss surface of a neural network with piecewise linear activations is partitioned into multiple smooth and multilinear cells by nondifferentiable boundaries. The constructed spurious local minima are concentrated in one cell as a valley: they are connected with each other by a continuous path, on which empirical risk is invariant. Further for one-hidden-layer networks, we prove that all local minima in a cell constitute an equivalence class; they are concentrated in a valley; and they are all global minima in the cell.
****
The problem with DDPG: understanding failures in deterministic environments with sparse rewards
Guillaume Matheron,Olivier Sigaud,Nicolas Perrin
In environments with continuous state and action spaces, state-of-the-art actor-critic reinforcement learning algorithms can solve very complex problems, yet can also fail in environments that seem trivial, but the reason for such failures is still poorly understood. In this paper, we contribute a formal explanation of these failures in the particular case of sparse reward and deterministic environments. First, using a very elementary control problem, we illustrate that the learning process can get
stuck into a fixed point corresponding to a poor solution. Then, generalizing from the studied example, we provide a detailed analysis of the underlying mechanisms which results in a new understanding of one of the convergence regimes of th

ese algorithms. The resulting perspective casts a new light on already existing solutions to the issues we have highlighted, and suggests other potential approaches.
****
LocalGAN: Modeling Local Distributions for Adversarial Response Generation
Zhen Xu,Baoxun Wang,Huan Zhang,Kexin Qiu,Deyuan Zhang,Chengjie Sun
This paper presents a new methodology for modeling the local semantic distribution of responses to a given query in the human-conversation corpus, and on this basis, explores a specified adversarial learning mechanism for training Neural Response Generation (NRG) models to build conversational agents. The proposed mechanism aims to address the training instability problem and improve the quality of generated results of Generative Adversarial Nets (GAN) in their utilizations in the response generation scenario. Our investigation begins with the thorough discussions upon the objective function brought by general GAN architectures to NRG models, and the training instability problem is proved to be ascribed to the special local distributions of conversational corpora. Consequently, an energy function is employed to estimate the status of a local area restricted by the query and its responses in the semantic space, and the mathematical approximation of this energy-based distribution is finally found. Building on this foundation, a local distribution oriented objective is proposed and combined with the original objective, working as a hybrid loss for the adversarial training of response generation models, named as LocalGAN. Our experimental results demonstrate that the reasonable local distribution modeling of the query-response corpus is of great importance to adversarial NRG, and our proposed LocalGAN is promising for improving both the training stability and the quality of generated results.

****
Generative Adversarial Networks For Data Scarcity Industrial Positron Images With Attention
Mingwei Zhu,Min Zhao,Min Yao,Ruipeng Guo
In the industrial field, the positron annihilation is not affected by complex environment, and the gamma-ray photon penetration is strong, so the nondestructive detection of industrial parts can be realized. Due to the poor image quality caused by gamma-ray photon scattering, attenuation and short sampling time in positron process, we propose the idea of combining deep learning to generate positron images with good quality and clear details by adversarial nets. The structure of the paper is as follows: firstly, we encode to get the hidden vectors of medical CT images based on transfer Learning, and use PCA to extract positron image features. Secondly, we construct a positron image memory based on attention mechanism as a whole input to the adversarial nets which uses medical hidden variables as a query. Finally, we train the whole model jointly and update the input parameters until convergence. Experiments have proved the possibility of generating rare positron images for industrial non-destructive testing using countermeasure networks, and good imaging results have been achieved.
****
OvA-INN: Continual Learning with Invertible Neural Networks
HOCQUET Guillaume,BICHLER Olivier,QUERLIOZ Damien
In the field of Continual Learning, the objective is to learn several tasks one after the other without access to the data from previous tasks. Several solutions have been proposed to tackle this problem but they usually assume that the user knows which of the tasks to perform at test time on a particular sample, or rely on small samples from previous data and most of them suffer of a substantial drop in accuracy when updated with batches of only one class at a time. In this article, we propose a new method, OvA-INN, which is able to learn one class at a time and without storing any of the previous data. To achieve this, for each class, we train a specific Invertible Neural Network to output the zero vector for its class. At test time, we can predict the class of a sample by identifying which network outputs the vector with the smallest norm. With this method, we show that we can take advantage of pretrained models by stacking an invertible network on top of a features extractor. This way, we are able to outperform state-of

-the-art approaches that rely on features learning for the Continual Learning of MNIST and CIFAR-100 datasets. In our experiments, we are reaching 72% accuracy on CIFAR-100 after training our model one class at a time.
****

Contextual Inverse Reinforcement Learning
Philip Korsunsky,Stav Belogolovsky,Tom Zahavy,Chen Tessler,Shie Mannor
We consider the Inverse Reinforcement Learning problem in Contextual Markov Decision Processes. In this setting, the reward, which is unknown to the agent, is a
function of a static parameter referred to as the context. There is also an “expert”
who knows this mapping and acts according to the optimal policy for each context.
The goal of the agent is to learn the expert’s mapping by observing demonstrations.
We define an optimization problem for finding this mapping and show that when
it is linear, the problem is convex. We present and analyze the sample complexity
of three algorithms for solving this problem: the mirrored descent algorithm,
evolution strategies, and the ellipsoid method. We also extend the first two methods
to work with general reward functions, e.g., deep neural networks, but without the
theoretical guarantees. Finally, we compare the different techniques empirically in
driving simulation and a medical treatment regime.
****

Learning Time-Aware Assistance Functions for Numerical Fluid Solvers
Kiwon Um,Yun (Raymond) Fei,Philipp Holl,Nils Thuerey
Improving the accuracy of numerical methods remains a central challenge in many disciplines and is especially important for nonlinear simulation problems. A representative example of such problems is fluid flow, which has been thoroughly studied to arrive at efficient simulations of complex flow phenomena. This paper presents a data-driven approach that learns to improve the accuracy of numerical solvers. The proposed method utilizes an advanced numerical scheme with a fine simulation resolution to acquire reference data. We, then, employ a neural network that infers a correction to move a coarse thus quickly obtainable result closer to the reference data. We provide insights into the targeted learning problem with different learning approaches: fully supervised learning methods with a naive and an optimized data acquisition as well as an unsupervised learning method with a differentiable Navier-Stokes solver. While our approach is very general and applicable to arbitrary partial differential equation models, we specifically highlight gains in accuracy for fluid flow simulations.
****

Transition Based Dependency Parser for Amharic Language Using Deep Learning
Mizanu Zelalem,Million Meshesha (PhD)
Researches shows that attempts done to apply existing dependency parser on morphological rich languages including Amharic shows a poor performance. In this study, a dependency parser for Amharic language is implemented using arc-eager transition system and LSTM network. The study introduced another way of building labeled dependency structure by using a separate network model to predict dependency relation. This helps the number of classes to decrease from 2n+2 into n, where n is the number of relationship types in the language and increases the number of examples for each class in the data set. Evaluation of the parser model results 91.54 and 81.4 unlabeled and labeled attachment score respectively. The major challenge in this study was the decrease of the accuracy of labeled attachment score. This is mainly due to the size and quality of the tree-bank available for Amharic language. Improving the tree-bank by increasing the size and by adding morphological information can make the performance of parser better.
****

Samples Are Useful? Not Always: denoising policy gradient updates using variance explained
Yannis Flet-Berliac,Philippe Preux
Policy gradient algorithms in reinforcement learning optimize the policy directly and rely on efficiently sampling an environment. However, while most sampling procedures are based solely on sampling the agent's policy, other measures directly accessible through these algorithms could be used to improve sampling before each policy update. Following this line of thoughts, we propose the use of SAUNA, a method where transitions are rejected from the gradient updates if they do not meet a particular criterion, and kept otherwise. This criterion, the fraction of variance explained Vex, is a measure of the discrepancy between a model and actual samples. In this work, Vex is used to evaluate the impact each transition will have on learning: this criterion refines sampling and improves the policy gradient algorithm. In this paper: (a) We introduce and explore Vex, the criterion used for denoising policy gradient updates. (b) We conduct experiments across a variety of benchmark environments, including standard continuous control problems. Our results show better performance with SAUNA. (c) We investigate why Vex provides a reliable assessment for the selection of samples that will positively impact learning. (d) We show how this criterion can work as a dynamic tool to adjust the ratio between exploration and exploitation.
****
Learning Surrogate Losses
Josif Grabocka,Randolf Scholz,Lars Schmidt-Thieme
The minimization of loss functions is the heart and soul of Machine Learning. In this paper, we propose an off-the-shelf optimization approach that can seamlessly minimize virtually any non-differentiable and non-decomposable loss function (e.g. Miss-classification Rate, AUC, F1, Jaccard Index, Mathew Correlation Coefficient, etc.). Our strategy learns smooth relaxation versions of the true losses by approximating them through a surrogate neural network. The proposed loss networks are set-wise models which are invariant to the order of mini-batch instances. Ultimately, the surrogate losses are learned jointly with the prediction model via bilevel optimization. Empirical results on multiple datasets with diverse real-life loss functions compared with state-of-the-art baselines demonstrate the efficiency of learning surrogate losses.
****
Boosting Network: Learn by Growing Filters and Layers via SplitLBI
Zuyuan Zhong,Chen Liu,Yanwei Fu,Yuan Yao
Network structures are important to learning good representations of many tasks in computer vision and machine learning communities. These structures are either manually designed, or searched by Neural Architecture Search (NAS) in previous works, which however requires either expert-level efforts, or prohibitive computational cost. In practice, it is desirable to efficiently and simultaneously learn both the structures and parameters of a network from arbitrary classes with budgeted computational cost. We identify it as a new learning paradigm -- Boosting Network, where one starts from simple models, delving into complex trained models progressively.
In this paper, by virtue of an iterative sparse regularization path -- Split Linearized Bregman Iteration (SplitLBI), we propose a simple yet effective boosting network method that can simultaneously grow and train a network by progressively adding both convolutional filters and layers. Extensive experiments with VGG and ResNets validate the effectiveness of our proposed algorithms.
****
Split LBI for Deep Learning: Structural Sparsity via Differential Inclusion Paths
Yanwei Fu,Chen Liu,Donghao Li,Xinwei Sun,Jinshan ZENG,Yuan Yao
Over-parameterization is ubiquitous nowadays in training neural networks to benefit both optimization in seeking global optima and generalization in reducing prediction error. However, compressive networks are desired in many real world applications and direct training of small networks may be trapped in local optima. In this paper, instead of pruning or distilling over-parameterized models to com

pressive ones, we propose a new approach based on \emph{differential inclusions of inverse scale spaces}, that generates a family of models from simple to complex ones by coupling gradient descent and mirror descent to explore model structural sparsity. It has a simple discretization, called the Split Linearized Bregman Iteration (SplitLBI), whose global convergence analysis in deep learning is established that from any initializations, algorithmic iterations converge to a critical point of empirical risks. Experimental evidence shows that\ SplitLBI may achieve state-of-the-art performance in large scale training on ImageNet-2012 dataset etc., while with \emph{early stopping} it unveils effective subnet architecture with comparable test accuracies to dense models after retraining instead of pruning well-trained ones.
****

Unsupervised Universal Self-Attention Network for Graph Classification
Dai Quoc Nguyen,Tu Dinh Nguyen,Dinh Phung
Existing graph embedding models often have weaknesses in exploiting graph structure similarities, potential dependencies among nodes and global network properties. To this end, we present U2GAN, a novel unsupervised model leveraging on the strength of the recently introduced universal self-attention network (Dehghani et al., 2019), to learn low-dimensional embeddings of graphs which can be used for graph classification. In particular, given an input graph, U2GAN first applies a self-attention computation, which is then followed by a recurrent transition to iteratively memorize its attention on vector representations of each node and its neighbors across each iteration. Thus, U2GAN can address the weaknesses in the existing models in order to produce plausible node embeddings whose sum is the final embedding of the whole graph. Experimental results show that our unsupervised U2GAN produces new state-of-the-art performances on a range of well-known benchmark datasets for the graph classification task. It even outperforms supervised methods in most of benchmark cases.
****

Manifold Modeling in Embedded Space: A Perspective for Interpreting "Deep Image Prior"
Tatsuya Yokota,Hidekata Hontani,Qibin Zhao,Andrzej Cichocki
Deep image prior (DIP), which utilizes a deep convolutional network (ConvNet) structure itself as an image prior, has attracted huge attentions in computer vision community. It empirically shows the effectiveness of ConvNet structure for various image restoration applications. However, why the DIP works so well is still unknown, and why convolution operation is essential for image reconstruction or enhancement is not very clear. In this study, we tackle these questions. The proposed approach is dividing the convolution into ``delay-embedding'' and ``transformation (\ie encoder-decoder)'', and proposing a simple, but essential, image/tensor modeling method which is closely related to dynamical systems and self-similarity. The proposed method named as manifold modeling in embedded space (MMES) is implemented by using a novel denoising-auto-encoder in combination with multi-way delay-embedding transform. In spite of its simplicity, the image/tensor completion and super-resolution results of MMES are quite similar even competitive to DIP in our extensive experiments, and these results would help us for re interpreting/characterizing the DIP from a perspective of ``low-dimensional patch-manifold prior''.
****

Novelty Detection Via Blurring
Sungik Choi,Sae-Young Chung
Conventional out-of-distribution (OOD) detection schemes based on variational autoencoder or Random Network Distillation (RND) are known to assign lower uncertainty to the OOD data than the target distribution. In this work, we discover that such conventional novelty detection schemes are also vulnerable to the blurred images. Based on the observation, we construct a novel RND-based OOD detector, SVD-RND, that utilizes blurred images during training. Our detector is simple, efficient in test time, and outperforms baseline OOD detectors in various domains. Further results show that SVD-RND learns a better target distribution representation than the baselines. Finally, SVD-RND combined with geometric transform a

chieves near-perfect detection accuracy in CelebA domain.
****
Small-GAN: Speeding up GAN Training using Core-Sets
Samarth Sinha,Han Zhang,Anirudh Goyal,Yoshua Bengio,Hugo Larochelle,Augustus Odena
BigGAN suggests that Generative Adversarial Networks (GANs) benefit disproportionately from large minibatch sizes. This finding is interesting but also discouraging -- large batch sizes are slow and expensive to emulate on conventional hardware. Thus, it would be nice if there were some trick by which we could generate batches that were effectively big though small in practice. In this work, we propose such a trick, inspired by the use of Coreset-selection in active learning. When training a GAN, we draw a large batch of samples from the prior and then compress that batch using Coreset-selection. To create effectively large batches of real images, we create a cached dataset of Inception activations of each training image, randomly project them down to a smaller dimension, and then use Coreset-selection on those projected embeddings at training time. We conduct experiments showing that this technique substantially reduces training time and memory usage for modern GAN variants, that it reduces the fraction of dropped modes in a synthetic dataset, and that it helps us use GANs to reach a new state of the art in anomaly detection.
****
Bounds on Over-Parameterization for Guaranteed Existence of Descent Paths in Shallow ReLU Networks
Arsalan Sharifnassab,Saber Salehkaleybar,S. Jamaloddin Golestani
We study the landscape of squared loss in neural networks with one-hidden layer and ReLU activation functions. Let m and d be the widths of hidden and input layers, respectively. We show that there exist poor local minima with positive curvature for some training sets of size $n\geq m+2d-2$. By positive curvature of a local minimum, we mean that within a small neighborhood the loss function is strictly increasing in all directions. Consequently, for such training sets, there are initialization of weights from which there is no descent path to global optima. It is known that for $n\le m$, there always exist descent paths to global optima from all initial weights. In this perspective, our results provide a somewhat sharp characterization of the over-parameterization required for "existence of descent paths" in the loss landscape.
****
Data-Independent Neural Pruning via Coresets
Ben Mussay,Margarita Osadchy,Vladimir Braverman,Samson Zhou,Dan Feldman
Previous work showed empirically that large neural networks can be significantly reduced in size while preserving their accuracy. Model compression became a central research topic, as it is crucial for deployment of neural networks on devices with limited computational and memory resources. The majority of the compression methods are based on heuristics and offer no worst-case guarantees on the trade-off between the compression rate and the approximation error for an arbitrarily new sample.

We propose the first efficient, data-independent neural pruning algorithm with a provable trade-off between its compression rate and the approximation error for any future test sample. Our method is based on the coreset framework, which finds a small weighted subset of points that provably approximates the original inputs. Specifically, we approximate the output of a layer of neurons by a coreset of neurons in the previous layer and discard the rest. We apply this framework in a layer-by-layer fashion from the top to the bottom. Unlike previous works, our coreset is data independent, meaning that it provably guarantees the accuracy of the function for any input $x\in \mathbb{R}^d$, including an adversarial one. We demonstrate the effectiveness of our method on popular network architectures. In particular, our coresets yield 90% compression of the LeNet-300-100 architecture on MNIST while improving the accuracy.
****
Deeper Insights into Weight Sharing in Neural Architecture Search

Yuge Zhang,Quanlu Zhang,Junyang Jiang,Zejun Lin,Yujing Wang
With the success of deep neural networks, Neural Architecture Search (NAS) as a way of automatic model design has attracted wide attention. As training every child model from scratch is very time-consuming, recent works leverage weight-sharing to speed up the model evaluation procedure. These approaches greatly reduce computation by maintaining a single copy of weights on the super-net and share the weights among every child model. However, weight-sharing has no theoretical guarantee and its impact has not been well studied before. In this paper, we conduct comprehensive experiments to reveal the impact of weight-sharing: (1) The best-performing models from different runs or even from consecutive epochs within the same run have significant variance; (2) Even with high variance, we can extract valuable information from training the super-net with shared weights; (3) The interference between child models is a main factor that induces high variance; (4) Properly reducing the degree of weight sharing could effectively reduce variance and improve performance.
****
Dirichlet Wrapper to Quantify Classification Uncertainty in Black-Box Systems
José Mena Roldán,Oriol Pujol Vila,Jordi Vitrià Marca
Nowadays, machine learning models are becoming a utility in many sectors. AI companies deliver pre-trained encapsulated models as application programming interfaces (APIs) that developers can combine with third party components, their models, and proprietary data, to create complex data products. This complexity and the lack of control and knowledge of the internals of these external components might cause unavoidable effects, such as lack of transparency, difficulty in auditability, and the emergence of uncontrolled potential risks. These issues are especially critical when practitioners use these components as black-boxes in new datasets. In order to provide actionable insights in this type of scenarios, in this work we propose the use of a wrapping deep learning model to enrich the output of a classification black-box with a measure of uncertainty. Given a black-box classifier, we propose a probabilistic neural network that works in parallel to the black-box and uses a Dirichlet layer as the fusion layer with the black-box. This Dirichlet layer yields a distribution on top of the multinomial output parameters of the classifier and enables the estimation of aleatoric uncertainty for any data sample.
Based on the resulting uncertainty measure, we advocate for a rejection system that selects the more confident predictions, discarding those more uncertain, leading to an improvement in the trustability of the resulting system. We showcase the proposed technique and methodology in two practical scenarios, one for NLP and another for computer vision, where a simulated API based is applied to different domains. Results demonstrate the effectiveness of the uncertainty computed by the wrapper and its high correlation to wrong predictions and misclassifications.
****
S2VG: Soft Stochastic Value Gradient method
Xiaoyu Tan,Chao Qu,Junwu Xiong,James Zhang
Model-based reinforcement learning (MBRL) has shown its advantages in sample-efficiency over model-free reinforcement learning (MFRL). Despite the impressive results it achieves, it still faces a trade-off between the ease of data generation and model bias. In this paper, we propose a simple and elegant model-based reinforcement learning algorithm called soft stochastic value gradient method (S2VG). S2VG combines the merits of the maximum-entropy reinforcement learning and MBRL, and exploits both real and imaginary data. In particular, we embed the model in the policy training and learn Q and V functions from the real (or imaginary) data set. Such embedding enables us to compute an analytic policy gradient through the back-propagation rather than the likelihood-ratio estimation, which can reduce the variance of the gradient estimation. We name our algorithm Soft Stochastic Value Gradient method to indicate its connection with the well-known stochastic value gradient method in \citep{heess2015Learning}.
****
Deep Network Classification by Scattering and Homotopy Dictionary Learning

John Zarka,Louis Thiry,Tomas Angles,Stephane Mallat
We introduce a sparse scattering deep convolutional neural network, which provid es a simple model to analyze properties of deep representation learning for clas sification. Learning a single dictionary matrix with a classifier yields a highe r classification accuracy than AlexNet over the ImageNet 2012 dataset. The netwo rk first applies a scattering transform that linearizes variabilities due to geo metric transformations such as translations and small deformations.
A sparse ℓ^1 dictionary coding reduces intra-class variability while preser ving class separation through projections over unions of linear spaces. It is im plemented in a deep convolutional network with a homotopy algorithm having an ex ponential convergence. A convergence proof is given in a general framework that includes ALISTA. Classification results are analyzed on ImageNet.
****
Scalable Generative Models for Graphs with Graph Attention Mechanism
Wataru Kawai,Yusuke Mukuta,Tatsuya Harada
Graphs are ubiquitous real-world data structures, and generative models that app roximate distributions over graphs and derive new samples from them have signifi cant importance. Among the known challenges in graph generation tasks, scalabili ty handling of large graphs and datasets is one of the most important for practi cal applications. Recently, an increasing number of graph generative models have been proposed and have demonstrated impressive results. However, scalability is still an unresolved problem due to the complex generation process or difficulty in training parallelization.
In this paper, we first define scalability from three different perspectives: nu mber of nodes, data, and node/edge labels. Then, we propose GRAM, a generative m odel for graphs that is scalable in all three contexts, especially in training. We aim to achieve scalability by employing a novel graph attention mechanism, fo rmulating the likelihood of graphs in a simple and general manner. Also, we appl y two techniques to reduce computational complexity. Furthermore, we construct a unified and non-domain-specific evaluation metric in node/edge-labeled graph ge neration tasks by combining a graph kernel and Maximum Mean Discrepancy. Our exp eriments on synthetic and real-world graphs demonstrated the scalability of our models and their superior performance compared with baseline methods.
****
Continuous Adaptation in Multi-agent Competitive Environments
Kuei-Tso Lee,Sheng-Jyh Wang
In a multi-agent competitive environment, we would expect an agent who can quick ly adapt to environmental changes may have a higher probability to survive and b eat other agents. In this paper, to discuss whether the adaptation capability ca n help a learning agent to improve its competitiveness in a multi-agent environm ent, we construct a simplified baseball game scenario to develop and evaluate th e adaptation capability of learning agents. Our baseball game scenario is modele d as a two-player zero-sum stochastic game with only the final reward. We purpos e a modified Deep CFR algorithm to learn a strategy that approximates the Nash e quilibrium strategy. We also form several teams, with different teams adopting d ifferent playing strategies, trying to analyze (1) whether an adaptation mechani sm can help in increasing the winning percentage and (2) what kind of initial st rategies can help a team to get a higher winning percentage. The experimental re sults show that the learned Nash-equilibrium strategy is very similar to real-li fe baseball game strategy. Besides, with the proposed strategy adaptation mechan ism, the winning percentage can be increased for the team with a Nash-equilibriu m initial strategy. Nevertheless, based on the same adaptation mechanism, those teams with deterministic initial strategies actually become less competitive.
****
Q-learning with UCB Exploration is Sample Efficient for Infinite-Horizon MDP
Yuanhao Wang,Kefan Dong,Xiaoyu Chen,Liwei Wang
A fundamental question in reinforcement learning is whether model-free algorithm s are sample efficient. Recently, Jin et al. (2018) proposed a Q-learning algor ithm with UCB exploration policy, and proved it has nearly optimal regret bound for finite-horizon episodic MDP. In this paper, we adapt Q-learning with UCB-exp

loration bonus to infinite-horizon MDP with discounted rewards \emph{without} accessing a generative model. We show that the \textit{sample complexity of exploration} of our algorithm is bounded by $\tilde{O}({\frac{SA}{\epsilon^2(1-\gamma)^7}})$. This improves the previously best known result of $\tilde{O}({\frac{SA}{\epsilon^4(1-\gamma)^8}})$ in this setting achieved by delayed Q-learning (Strehlet al., 2006),, and matches the lower bound in terms of ϵ as well as S and A up to logarithmic factors.
****
Robust Cross-lingual Embeddings from Parallel Sentences
Ali Sabet,Prakhar Gupta,Jean-Baptiste Cordonnier,Robert West,Martin Jaggi
Recent advances in cross-lingual word embeddings have primarily relied on mapping-based methods, which project pretrained word embeddings from different languages into a shared space through a linear transformation. However, these approaches assume word embedding spaces are isomorphic between different languages, which has been shown not to hold in practice (Søgaard et al., 2018), and fundamentally limits their performance. This motivates investigating joint learning methods which can overcome this impediment, by simultaneously learning embeddings across languages via a cross-lingual term in the training objective. Given the abundance of parallel data available (Tiedemann, 2012), we propose a bilingual extension of the CBOW method which leverages sentence-aligned corpora to obtain robust cross-lingual word and sentence representations. Our approach significantly improves cross-lingual sentence retrieval performance over all other approaches, as well as convincingly outscores mapping methods while maintaining parity with jointly trained methods on word-translation. It also achieves parity with a deep RNN method on a zero-shot cross-lingual document classification task, requiring far fewer computational resources for training and inference. As an additional advantage, our bilingual method also improves the quality of monolingual word vectors despite training on much smaller datasets. We make our code and models publicly available.

****
Semi-supervised Learning by Coaching
Hieu Pham,Quoc V. Le
Recent semi-supervised learning (SSL) methods often have a teacher to train a student in order to propagate labels from labeled data to unlabeled data. We argue that a weakness of these methods is that the teacher does not learn from the student's mistakes during the course of student's learning. To address this weakness, we introduce Coaching, a framework where a teacher generates pseudo labels for unlabeled data, from which a student will learn and the student's performance on labeled data will be used as reward to train the teacher using policy gradient.

Our experiments show that Coaching significantly improves over state-of-the-art SSL baselines. For instance, on CIFAR-10, with only 4,000 labeled examples, a WideResNet-28-2 trained by Coaching achieves 96.11% accuracy, which is better than 94.9% achieved by the same architecture trained with 45,000 labeled. On ImageNet with 10% labeled examples, Coaching trains a ResNet-50 to 72.94% top-1 accuracy, comfortably outperforming the existing state-of-the-art by more than 4%. Coaching also scales successfully to the high data regime with full ImageNet. Specifically, with additional 9 million unlabeled images from OpenImages, Coaching trains a ResNet-50 to 82.34% top-1 accuracy, setting a new state-of-the-art for the architecture on ImageNet without using extra labeled data.
****
DYNAMIC SELF-TRAINING FRAMEWORK FOR GRAPH CONVOLUTIONAL NETWORKS
Ziang Zhou,Shenzhong Zhang,Zengfeng Huang
Graph neural networks (GNN) such as GCN, GAT, MoNet have achieved state-of-the-art results on semi-supervised learning on graphs. However, when the number of labeled nodes is very small, the performances of GNNs downgrade dramatically. Self-training has proved to be effective for resolving this issue, however, the performance of self-trained GCN is still inferior to that of G2G and DGI for many se

ttings. Moreover, additional model complexity make it more difficult to tune the hyper-parameters and do model selection. We argue that the power of self-training is still not fully explored for the node classification task. In this paper, we propose a unified end-to-end self-training framework called \emph{Dynamic Self-traning}, which generalizes and simplifies prior work. A simple instantiation of the framework based on GCN is provided and empirical results show that our framework outperforms all previous methods including GNNs, embedding based method and self-trained GCNs by a noticeable margin. Moreover, compared with standard self-training, hyper-parameter tuning for our framework is easier.
****
Blockwise Self-Attention for Long Document Understanding
Jiezhong Qiu,Hao Ma,Omer Levy,Scott Wen-tau Yih,Sinong Wang,Jie Tang
We present BlockBERT, a lightweight and efficient BERT model that is designed to better modeling long-distance dependencies. Our model extends BERT by introducing sparse block structures into the attention matrix to reduce both memory consumption and training time, which also enables attention heads to capture either short- or long-range contextual information. We conduct experiments on several benchmark question answering datasets with various paragraph lengths. Results show that BlockBERT uses 18.7-36.1% less memory and reduces the training time by 12.0-25.1%, while having comparable and sometimes better prediction accuracy, compared to an advanced BERT-based model, RoBERTa.
****
Mixout: Effective Regularization to Finetune Large-scale Pretrained Language Models
Cheolhyoung Lee,Kyunghyun Cho,Wanmo Kang
In natural language processing, it has been observed recently that generalization could be greatly improved by finetuning a large-scale language model pretrained on a large unlabeled corpus. Despite its recent success and wide adoption, finetuning a large pretrained language model on a downstream task is prone to degenerate performance when there are only a small number of training instances available. In this paper, we introduce a new regularization technique, to which we refer as “mixout”, motivated by dropout. Mixout stochastically mixes the parameters of two models. We show that our mixout technique regularizes learning to minimize the deviation from one of the two models and that the strength of regularization adapts along the optimization trajectory. We empirically evaluate the proposed mixout and its variants on finetuning a pretrained language model on downstream tasks. More specifically, we demonstrate that the stability of finetuning and the average accuracy greatly increase when we use the proposed approach to regularize finetuning of BERT on downstream tasks in GLUE.
****
I Am Going MAD: Maximum Discrepancy Competition for Comparing Classifiers Adaptively
Haotao Wang,Tianlong Chen,Zhangyang Wang,Kede Ma
The learning of hierarchical representations for image classification has experienced an impressive series of successes due in part to the availability of large-scale labeled data for training. On the other hand, the trained classifiers have traditionally been evaluated on small and fixed sets of test images, which are deemed to be extremely sparsely distributed in the space of all natural images. It is thus questionable whether recent performance improvements on the excessively re-used test sets generalize to real-world natural images with much richer content variations. Inspired by efficient stimulus selection for testing perceptual models in psychophysical and physiological studies, we present an alternative framework for comparing image classifiers, which we name the MAximum Discrepancy (MAD) competition. Rather than comparing image classifiers using fixed test images, we adaptively sample a small test set from an arbitrarily large corpus of unlabeled images so as to maximize the discrepancies between the classifiers, measured by the distance over WordNet hierarchy. Human labeling on the resulting model-dependent image sets reveals the relative performance of the competing classifiers, and provides useful insights on potential ways to improve them. We report the MAD competition results of eleven ImageNet classifiers while noting that

the framework is readily extensible and cost-effective to add future classifiers into the competition. Codes can be found at https://github.com/TAMU-VITA/MAD.
****
Black-Box Adversarial Attack with Transferable Model-based Embedding
Zhichao Huang,Tong Zhang
We present a new method for black-box adversarial attack. Unlike previous methods that combined transfer-based and scored-based methods by using the gradient or initialization of a surrogate white-box model, this new method tries to learn a low-dimensional embedding using a pretrained model, and then performs efficient search within the embedding space to attack an unknown target network. The method produces adversarial perturbations with high level semantic patterns that are easily transferable. We show that this approach can greatly improve the query efficiency of black-box adversarial attack across different target network architectures. We evaluate our approach on MNIST, ImageNet and Google Cloud Vision API, resulting in a significant reduction on the number of queries. We also attack adversarially defended networks on CIFAR10 and ImageNet, where our method not only reduces the number of queries, but also improves the attack success rate.
****
Stabilizing Off-Policy Reinforcement Learning with Conservative Policy Gradients
Chen Tessler,Nadav Merlis,Shie Mannor
In recent years, advances in deep learning have enabled the application of reinforcement learning algorithms in complex domains. However, they lack the theoretical guarantees which are present in the tabular setting and suffer from many stability and reproducibility problems \citep{henderson2018deep}. In this work, we suggest a simple approach for improving stability and providing probabilistic performance guarantees in off-policy actor-critic deep reinforcement learning regimes. Experiments on continuous action spaces, in the MuJoCo control suite, show that our proposed method reduces the variance of the process and improves the overall performance.
****
Do Image Classifiers Generalize Across Time?
Vaishaal Shankar,Achal Dave,Rebecca Roelofs,Deva Ramanan,Ben Recht,Ludwig Schmidt
We study the robustness of image classifiers to temporal perturbations derived from videos. As part of this study, we construct ImageNet-Vid-Robust and YTBB-Robust, containing a total 57,897 images grouped into 3,139 sets of perceptually similar images. Our datasets were derived from ImageNet-Vid and Youtube-BB respectively and thoroughly re-annotated by human experts for image similarity. We evaluate a diverse array of classifiers pre-trained on ImageNet and show a median classification accuracy drop of 16 and 10 percent on our two datasets. Additionally, we evaluate three detection models and show that natural perturbations induce both classification as well as localization errors, leading to a median drop in detection mAP of 14 points. Our analysis demonstrates that perturbations occurring naturally in videos pose a substantial and realistic challenge to deploying convolutional neural networks in environments that require both reliable and low-latency predictions.
****
Cross-Domain Few-Shot Classification via Learned Feature-Wise Transformation
Hung-Yu Tseng,Hsin-Ying Lee,Jia-Bin Huang,Ming-Hsuan Yang
Few-shot classification aims to recognize novel categories with only few labeled images in each class. Existing metric-based few-shot classification algorithms predict categories by comparing the feature embeddings of query images with those from a few labeled images (support examples) using a learned metric function. While promising performance has been demonstrated, these methods often fail to generalize to unseen domains due to large discrepancy of the feature distribution across domains. In this work, we address the problem of few-shot classification under domain shifts for metric-based methods. Our core idea is to use feature-wise transformation layers for augmenting the image features using affine transforms to simulate various feature distributions under different domains in the training stage. To capture variations of the feature distributions under different

domains, we further apply a learning-to-learn approach to search for the hyper-parameters of the feature-wise transformation layers. We conduct extensive experiments and ablation studies under the domain generalization setting using five few-shot classification datasets: mini-ImageNet, CUB, Cars, Places, and Plantae. Experimental results demonstrate that the proposed feature-wise transformation layer is applicable to various metric-based models, and provides consistent improvements on the few-shot classification performance under domain shift.
****
Evolutionary Reinforcement Learning for Sample-Efficient Multiagent Coordination
Shauharda Khadka,Somdeb Majumdar,Santiago Miret,Stephen McAleer,Kagan Tumer
Many cooperative multiagent reinforcement learning environments provide agents with a sparse team-based reward as well as a dense agent-specific reward that incentivizes learning basic skills. Training policies solely on the team-based reward is often difficult due to its sparsity. Also, relying solely on the agent-specific reward is sub-optimal because it usually does not capture the team coordination objective. A common approach is to use reward shaping to construct a proxy reward by combining the individual rewards. However, this requires manual tuning for each environment. We introduce Multiagent Evolutionary Reinforcement Learning (MERL), a split-level training platform that handles the two objectives separately through two optimization processes. An evolutionary algorithm maximizes the sparse team-based objective through neuroevolution on a population of teams. Concurrently, a gradient-based optimizer trains policies to only maximize the dense agent-specific rewards. The gradient-based policies are periodically added to the evolutionary population as a way of information transfer between the two optimization processes. This enables the evolutionary algorithm to use skills learned via the agent-specific rewards toward optimizing the global objective. Results demonstrate that MERL significantly outperforms state-of-the-art methods such as MADDPG on a number of difficult coordination benchmarks.
****
A shallow feature extraction network with a large receptive field for stereo matching tasks
Jianguo Liu,Yunjian Feng,Guo Ji,Fuwu Yan
Stereo matching is one of the important basic tasks in the computer vision field. In recent years, stereo matching algorithms based on deep learning have achieved excellent performance and become the mainstream research direction. Existing algorithms generally use deep convolutional neural networks (DCNNs) to extract more abstract semantic information, but we believe that the detailed information of the spatial structure is more important for stereo matching tasks. Based on this point of view, this paper proposes a shallow feature extraction network with a large receptive field. The network consists of three parts: a primary feature extraction module, an atrous spatial pyramid pooling (ASPP) module and a feature fusion module. The primary feature extraction network contains only three convolution layers. This network utilizes the basic feature extraction ability of the shallow network to extract and retain the detailed information of the spatial structure. In this paper, the dilated convolution and atrous spatial pyramid pooling (ASPP) module is introduced to increase the size of receptive field. In addition, a feature fusion module is designed, which integrates the feature maps with multiscale receptive fields and mutually complements the feature information of different scales. We replaced the feature extraction part of the existing stereo matching algorithms with our shallow feature extraction network, and achieved state-of-the-art performance on the KITTI 2015 dataset. Compared with the reference network, the number of parameters is reduced by 42%, and the matching accuracy is improved by 1.9%.
****
Learning Boolean Circuits with Neural Networks
Eran Malach,Shai Shalev-Shwartz
Training neural-networks is computationally hard. However, in practice they are trained efficiently using gradient-based algorithms, achieving remarkable performance on natural data. To bridge this gap, we observe the property of local correlation: correlation between small patterns of the input and the target label. W

e focus on learning deep neural-networks with a variant of gradient-descent, whe
n the target function is a tree-structured Boolean circuit. We show that in this
 case, the existence of correlation between the gates of the circuit and the tar
get label determines whether the optimization succeeds or fails. Using this resu
lt, we show that neural-networks can learn the (log n)-parity problem for most p
roduct distributions. These results hint that local correlation may play an impo
rtant role in differentiating between distributions that are hard or easy to lea
rn.
****
Towards Principled Objectives for Contrastive Disentanglement
Anwesa Choudhuri,Ashok Vardhan Makkuva,Ranvir Rana,Sewoong Oh,Girish Chowdhary,A
lexander Schwing
Unsupervised learning is an important tool that has received a significant amoun
t of attention for decades. Its goal is `unsupervised recovery,' i.e., extractin
g salient factors/properties from unlabeled data. Because of the challenges in
defining salient properties, recently, `contrastive disentanglement' has gained
popularity to discover the additional variations that are enhanced in one datase
t relative to another. %In fact, contrastive disentanglement and unsupervised re
covery are often combined in that we seek additional variations that exhibit sal
ient factors/properties.
Existing formulations have devised a variety of losses for this task. However, a
ll present day methods exhibit two major shortcomings: (1) encodings for data th
at does not exhibit salient factors is not pushed to carry no signal; and (2) in
troduced losses are often hard to estimate and require additional trainable para
meters. We present a new formulation for contrastive disentanglement which avoid
s both shortcomings by carefully formulating a probabilistic model and by using
non-parametric yet easily computable metrics. We show on four challenging datase
ts that the proposed approach is able to better disentangle salient factors.

****
Compositional languages emerge in a neural iterated learning model
Yi Ren,Shangmin Guo,Matthieu Labeau,Shay B. Cohen,Simon Kirby
The principle of compositionality, which enables natural language to represent c
omplex concepts via a structured combination of simpler ones, allows us to conve
y an open-ended set of messages using a limited vocabulary. If compositionality
is indeed a natural property of language, we may expect it to appear in communic
ation protocols that are created by neural agents via grounded language learning
. Inspired by the iterated learning framework, which simulates the process of la
nguage evolution, we propose an effective neural iterated learning algorithm tha
t, when applied to interacting neural agents, facilitates the emergence of a mor
e structured type of language. Indeed, these languages provide specific advantag
es to neural agents during training, which translates as a larger posterior prob
ability, which is then incrementally amplified via the iterated learning procedu
re. Our experiments confirm our analysis, and also demonstrate that the emerged
languages largely improve the generalization of the neural agent communication.
****
Population-Guided Parallel Policy Search for Reinforcement Learning
Whiyoung Jung,Giseung Park,Youngchul Sung
In this paper, a new population-guided parallel learning scheme is proposed to e
nhance the performance of off-policy reinforcement learning (RL). In the propose
d scheme, multiple identical learners with their own value-functions and policie
s share a common experience replay buffer, and search a good policy in collabora
tion with the guidance of the best policy information. The key point is that the
 information of the best policy is fused in a soft manner by constructing an au
gmented loss function for policy update to enlarge the overall search region by
the multiple learners. The guidance by the previous best policy and the enlarged
  range enable faster and better policy search, and monotone improvement of the
expected cumulative return by the proposed scheme is proved theoretically. Worki
ng algorithms are constructed by applying the proposed scheme to the twin delaye
d deep deterministic (TD3) policy gradient algorithm, and numerical results show

that the constructed P3S-TD3 outperforms most of the current state-of-the-art RL algorithms, and the gain is significant in the case of sparse reward environment.
****
Variational Recurrent Models for Solving Partially Observable Control Tasks
Dongqi Han,Kenji Doya,Jun Tani
In partially observable (PO) environments, deep reinforcement learning (RL) agents often suffer from unsatisfactory performance, since two problems need to be tackled together: how to extract information from the raw observations to solve the task, and how to improve the policy. In this study, we propose an RL algorithm for solving PO tasks. Our method comprises two parts: a variational recurrent model (VRM) for modeling the environment, and an RL controller that has access to both the environment and the VRM. The proposed algorithm was tested in two types of PO robotic control tasks, those in which either coordinates or velocities were not observable and those that require long-term memorization. Our experiments show that the proposed algorithm achieved better data efficiency and/or learned more optimal policy than other alternative approaches in tasks in which unobserved states cannot be inferred from raw observations in a simple manner.
****
Learning to Discretize: Solving 1D Scalar Conservation Laws via Deep Reinforcement Learning
Yufei Wang*,Ziju Shen*,Zichao Long,Bin Dong
Conservation laws are considered to be fundamental laws of nature. It has broad application in many fields including physics, chemistry, biology, geology, and engineering. Solving the differential equations associated with conservation laws is a major branch in computational mathematics. Recent success of machine learning, especially deep learning, in areas such as computer vision and natural language processing, has attracted a lot of attention from the community of computational mathematics and inspired many intriguing works in combining machine learning with traditional methods. In this paper, we are the first to explore the possibility and benefit of solving nonlinear conservation laws using deep reinforcement learning. As a proof of concept, we focus on 1-dimensional scalar conservation laws. We deploy the machinery of deep reinforcement learning to train a policy network that can decide on how the numerical solutions should be approximated in a sequential and spatial-temporal adaptive manner. We will show that the problem of solving conservation laws can be naturally viewed as a sequential decision making process and the numerical schemes learned in such a way can easily enforce long-term accuracy.
Furthermore, the learned policy network is carefully designed to determine a good local discrete approximation based on the current state of the solution, which essentially makes the proposed method a meta-learning approach.
In other words, the proposed method is capable of learning how to discretize for a given situation mimicking human experts. Finally, we will provide details on how the policy network is trained, how well it performs compared with some state-of-the-art numerical solvers such as WENO schemes, and how well it generalizes. Our code is released anomynously at \url{https://github.com/qwerlanksdf/L2D}.
****
Composable Semi-parametric Modelling for Long-range Motion Generation
Jingwei Xu,Huazhe Xu,Bingbing Ni,Xiaokang Yang,Trevor Darrell
Learning diverse and natural behaviors is one of the longstanding goal for creating intelligent characters in the animated world. In this paper, we propose ``Composable Semi-parametric MOdelling'' (COSMO), a method for generating long range diverse and distinctive behaviors to achieve a specific goal location. Our proposed method learns to model the motion of human by combining the complementary strengths of both non-parametric techniques and parametric ones. Given the starting and ending state, a memory bank is used to retrieve motion references that are provided as source material to a deep network. The synthesis is performed by a deep network that controls the style of the provided motion material and modifies it to become natural. On skeleton datasets with diverse motion, we show that the proposed method outperforms existing parametric and non-parametric baselines

. We also demonstrate the generated sequences are useful as subgoals for actual physical execution in the animated world.
****

Towards an Adversarially Robust Normalization Approach
Muhammad Awais,Fahad Shamshad,Sung-Ho Bae
Batch Normalization (BatchNorm) has shown to be effective for improving and accelerating the training of deep neural networks. However, recently it has been shown that it is also vulnerable to adversarial perturbations. In this work, we aim to investigate the cause of adversarial vulnerability of the BatchNorm. We hypothesize that the use of different normalization statistics during training and inference (mini-batch statistics for training and moving average of these values at inference) is the main cause of this adversarial vulnerability in the BatchNorm layer. We empirically proved this by experiments on various neural network architectures and datasets. Furthermore, we introduce Robust Normalization (RobustNorm) and experimentally show that it is not only resilient to adversarial perturbation but also inherit the benefits of BatchNorm.
****

Generative Latent Flow
Zhisheng Xiao,Qing Yan,Yali Amit
In this work, we propose the Generative Latent Flow (GLF), an algorithm for generative modeling of the data distribution. GLF uses an Auto-encoder (AE) to learn latent representations of the data, and a normalizing flow to map the distribution of the latent variables to that of simple i.i.d noise. In contrast to some other Auto-encoder based generative models, which use various regularizers that encourage the encoded latent distribution to match the prior distribution, our model explicitly constructs a mapping between these two distributions, leading to better density matching while avoiding over regularizing the latent variables. We compare our model with several related techniques, and show that it has many relative advantages including fast convergence, single stage training and minimal reconstruction trade-off. We also study the relationship between our model and its stochastic counterpart, and show that our model can be viewed as a vanishing noise limit of VAEs with flow prior. Quantitatively, under standardized evaluations, our method achieves state-of-the-art sample quality and diversity among AE based models on commonly used datasets, and is competitive with GANs' benchmarks.
****

GAT: Generative Adversarial Training for Adversarial Example Detection and Robust Classification
Xuwang Yin,Soheil Kolouri,Gustavo K Rohde
The vulnerabilities of deep neural networks against adversarial examples have become a significant concern for deploying these models in sensitive domains. Devising a definitive defense against such attacks is proven to be challenging, and the methods relying on detecting adversarial samples are only valid when the attacker is oblivious to the detection mechanism. In this paper we propose a principled adversarial example detection method that can withstand norm-constrained white-box attacks. Inspired by one-versus-the-rest classification, in a K class classification problem, we train K binary classifiers where the i-th binary classifier is used to distinguish between clean data of class i and adversarially perturbed samples of other classes. At test time, we first use a trained classifier to get the predicted label (say k) of the input, and then use the k-th binary classifier to determine whether the input is a clean sample (of class k) or an adversarially perturbed example (of other classes). We further devise a generative approach to detecting/classifying adversarial examples by interpreting each binary classifier as an unnormalized density model of the class-conditional data. We provide comprehensive evaluation of the above adversarial example detection/classification methods, and demonstrate their competitive performances and compelling properties. Code is available at https://github.com/xuwangyin/GAT-Generative-Adversarial-Training
****

CZ-GEM: A FRAMEWORK FOR DISENTANGLED REPRESENTATION LEARNING

Akash Srivastava,Yamini Bansal,Yukun Ding,Bernhard Egger,Prasanna Sattigeri,Josh Tenenbaum,David D. Cox,Dan Gutfreund
Learning disentangled representations of data is one of the central themes in unsupervised learning in general and generative modelling in particular. In this work, we tackle a slightly more intricate scenario where the observations are generated from a conditional distribution of some known control variate and some latent noise variate. To this end, we present a hierarchical model and a training method (CZ-GEM) that leverages some of the recent developments in likelihood-based and likelihood-free generative models. We show that by formulation, CZ-GEM introduces the right inductive biases that ensure the disentanglement of the control from the noise variables, while also keeping the components of the control variate disentangled. This is achieved without compromising on the quality of the generated samples. Our approach is simple, general, and can be applied both in supervised and unsupervised settings.
****

Generalized Natural Language Grounded Navigation via Environment-agnostic Multitask Learning
Xin Wang,Vihan Jain,Eugene Ie,William Wang,Zornitsa Kozareva,Sujith Ravi
Recent research efforts enable study for natural language grounded navigation in photo-realistic environments, e.g., following natural language instructions or dialog. However, existing methods tend to overfit training data in seen environments and fail to generalize well in previously unseen environments. In order to close the gap between seen and unseen environments, we aim at learning a generalizable navigation model from two novel perspectives:
(1) we introduce a multitask navigation model that can be seamlessly trained on both Vision-Language Navigation (VLN) and Navigation from Dialog History (NDH) tasks, which benefits from richer natural language guidance and effectively transfers knowledge across tasks;
(2) we propose to learn environment-agnostic representations for navigation policy that are invariant among environments, thus generalizing better on unseen environments.
Extensive experiments show that our environment-agnostic multitask navigation model significantly reduces the performance gap between seen and unseen environments and outperforms the baselines on unseen environments by 16% (relative measure on success rate) on VLN and 120% (goal progress) on NDH, establishing the new state of the art for NDH task.
****

Global Concavity and Optimization in a Class of Dynamic Discrete Choice Models
Yiding Feng,Ekaterina Khmelnitskaya,Denis Nekipelov
Discrete choice models with unobserved heterogeneity are commonly used Econometric models for dynamic Economic behavior which have been adopted in practice to predict behavior of individuals and firms from schooling and job choices to strategic decisions in market competition. These models feature optimizing agents who choose among a finite set of options in a sequence of periods and receive choice-specific payoffs that depend on both variables that are observed by the agent and recorded in the data and variables that are only observed by the agent but not recorded in the data. Existing work in Econometrics assumes that optimizing agents are fully rational and requires finding a functional fixed point to find the optimal policy. We show that in an important class of discrete choice models the value function is globally concave in the policy. That means that simple algorithms that do not require fixed point computation, such as the policy gradient algorithm, globally converge to the optimal policy. This finding can both be used to relax behavioral assumption regarding the optimizing agents and to facilitate Econometric analysis of dynamic behavior. In particular, we demonstrate significant computational advantages in using a simple implementation policy gradient algorithm over existing "nested fixed point" algorithms used in Econometrics.
****

Posterior sampling for multi-agent reinforcement learning: solving extensive games with imperfect information
Yichi Zhou,Jialian Li,Jun Zhu

Posterior sampling for reinforcement learning (PSRL) is a useful framework for m
aking decisions in an unknown environment. PSRL maintains a posterior distribut
ion of the environment and then makes planning on the environment sampled from t
he posterior distribution. Though PSRL works well on single-agent reinforcement
 learning problems, how to apply PSRL to multi-agent reinforcement learning probl
ems is relatively unexplored. In this work, we extend PSRL to two-player zero-su
m extensive-games with imperfect information (TEGI), which is a class of multi-a
gent systems. More specifically, we combine PSRL with counterfactual regret mini
mization (CFR), which is the leading algorithm for TEGI with a known environment
. Our main contribution is a novel design of interaction strategies. With our in
teraction strategies, our algorithm provably converges to the Nash Equilibrium a
t a rate of $O(\sqrt{\log T/T})$. Empirical results show that our algorithm work
s well.
****
On the Pareto Efficiency of Quantized CNN
Ting-Wu Chin,Pierce I-Jen Chuang,Vikas Chandra,Diana Marculescu
Weight Quantization for deep convolutional neural networks (CNNs) has shown prom
ising results in compressing and accelerating CNN-powered applications such as s
emantic segmentation, gesture recognition, and scene understanding. Prior art ha
s shown that different datasets, tasks, and network architectures admit differen
t iso-accurate precision values, which increase the complexity of efficient quan
tized neural network implementations from both hardware and software perspective
s. In this work, we show that when the number of channels is allowed to vary in
 an iso-model size scenario, lower precision values Pareto dominate higher precis
ion ones (in accuracy vs. model size) for networks with standard convolutions. R
elying on comprehensive empirical analyses, we find that the Pareto optimal prec
ision value of a convolution layer depends on the number of input channels per o
utput filters and provide theoretical insights for it. To this end, we develop a
 simple algorithm to select the precision values for CNNs that outperforms corre
sponding 8-bit quantized networks by 0.9% and 2.2% in top-1 accuracy on ImageNet
 for ResNet50 and MobileNetV2, respectively.
****
BANANAS: Bayesian Optimization with Neural Networks for Neural Architecture Sear
ch
Colin White,Willie Neiswanger,Yash Savani
Neural Architecture Search (NAS) has seen an explosion of research in the past f
ew years. A variety of methods have been proposed to perform NAS, including rein
forcement learning, Bayesian optimization with a Gaussian process model, evoluti
onary search, and gradient descent. In this work, we design a NAS algorithm that
 performs Bayesian optimization using a neural network model.

We develop a path-based encoding scheme to featurize the neural architectures th
at are used to train the neural network model. This strategy is particularly eff
ective for encoding architectures in cell-based search spaces. After training on
 just 200 random neural architectures, we are able to predict the validation acc
uracy of a new architecture to within one percent of its true accuracy on averag
e. This may be of independent interest beyond Bayesian neural architecture searc
h.

We test our algorithm on the NASBench dataset (Ying et al. 2019), and show that
our algorithm significantly outperforms other NAS methods including evolutionary
 search, reinforcement learning, and AlphaX (Wang et al. 2019). Our algorithm is
 over 100x more efficient than random search, and 3.8x more efficient than the n
ext-best algorithm. We also test our algorithm on the search space used in DARTS
 (Liu et al. 2018), and show that our algorithm is competitive with state-of-the
-art NAS algorithms on this search space.
****
Potential Flow Generator with L_2 Optimal Transport Regularity for Generative
Models
Liu Yang,George Em Karniadakis

We propose a potential flow generator with L_2 optimal transport regularity, which can be easily integrated into a wide range of generative models including different versions of GANs and flow-based models. With up to a slight augmentation of the original generator loss functions, our generator is not only a transport map from the input distribution to the target one, but also the one with minimum L_2 transport cost. We show the correctness and robustness of the potential flow generator in several 2D problems, and illustrate the concept of ``proximity'' due to the L_2 optimal transport regularity. Subsequently, we demonstrate the effectiveness of the potential flow generator in image translation tasks with unpaired training data from the MNIST dataset and the CelebA dataset.
****

Integrative Tensor-based Anomaly Detection System For Satellites
Youjin Shin,Sangyup Lee,Shahroz Tariq,Myeong Shin Lee,OkchulJung,Daewon Chung,Simon Woo
Detecting anomalies is of growing importance for various industrial applications and mission-critical infrastructures, including satellite systems. Although there have been several studies in detecting anomalies based on rule-based or machine learning-based approaches for satellite systems, a tensor-based decomposition method has not been extensively explored for anomaly detection. In this work, we introduce an Integrative Tensor-based Anomaly Detection (ITAD) framework to detect anomalies in a satellite system. Because of the high risk and cost, detecting anomalies in a satellite system is crucial. We construct 3rd-order tensors with telemetry data collected from Korea Multi-Purpose Satellite-2 (KOMPSAT-2) and calculate the anomaly score using one of the component matrices obtained by applying CANDECOMP/PARAFAC decomposition to detect anomalies. Our result shows that our tensor-based approach can be effective in achieving higher accuracy and reducing false positives in detecting anomalies as compared to other existing approaches.
****

Detecting and Diagnosing Adversarial Images with Class-Conditional Capsule Reconstructions
Yao Qin,Nicholas Frosst,Sara Sabour,Colin Raffel,Garrison Cottrell,Geoffrey Hinton
Adversarial examples raise questions about whether neural network models are sensitive to the same visual features as humans. In this paper, we first detect adversarial examples or otherwise corrupted images based on a class-conditional reconstruction of the input. To specifically attack our detection mechanism, we propose the Reconstructive Attack which seeks both to cause a misclassification and a low reconstruction error. This reconstructive attack produces undetected adversarial examples but with much smaller success rate. Among all these attacks, we find that CapsNets always perform better than convolutional networks. Then, we diagnose the adversarial examples for CapsNets and find that the success of the reconstructive attack is highly related to the visual similarity between the source and target class. Additionally, the resulting perturbations can cause the input image to appear visually more like the target class and hence become non-adversarial. This suggests that CapsNets use features that are more aligned with human perception and have the potential to address the central issue raised by adversarial examples.
****

MACER: Attack-free and Scalable Robust Training via Maximizing Certified Radius
Runtian Zhai,Chen Dan,Di He,Huan Zhang,Boqing Gong,Pradeep Ravikumar,Cho-Jui Hsieh,Liwei Wang
Adversarial training is one of the most popular ways to learn robust models but is usually attack-dependent and time costly. In this paper, we propose the MACER algorithm, which learns robust models without using adversarial training but performs better than all existing provable l2-defenses. Recent work shows that randomized smoothing can be used to provide a certified l2 radius to smoothed classifiers, and our algorithm trains provably robust smoothed classifiers via MAximizing the CErtified Radius (MACER). The attack-free characteristic makes MACER faster to train and easier to optimize. In our experiments, we show that our metho

d can be applied to modern deep neural networks on a wide range of datasets, inc luding Cifar-10, ImageNet, MNIST, and SVHN. For all tasks, MACER spends less tra ining time than state-of-the-art adversarial training algorithms, and the learne d models achieve larger average certified radius.
****
TinyBERT: Distilling BERT for Natural Language Understanding
Xiaoqi Jiao,Yichun Yin,Lifeng Shang,Xin Jiang,Xiao Chen,Linlin Li,Fang Wang,Qun Liu
Language model pre-training, such as BERT, has significantly improved the perfor mances of many natural language processing tasks. However, the pre-trained langu age models are usually computationally expensive and memory intensive, so it is difficult to effectively execute them on resource-restricted devices. To acceler ate inference and reduce model size while maintaining accuracy, we firstly propo se a novel Transformer distillation method that is specially designed for knowle dge distillation (KD) of the Transformer-based models. By leveraging this new KD method, the plenty of knowledge encoded in a large “teacher” BERT can be well t ransferred to a small “student” TinyBERT. Moreover, we introduce a new two-stage learning framework for TinyBERT, which performs Transformer distillation at bot h the pre-training and task-specific learning stages. This framework ensures tha t TinyBERT can capture the general domain as well as the task-specific knowledge in BERT. TinyBERT is empirically effective and achieves comparable results with BERT on GLUE benchmark, while being 7.5x smaller and 9.4x faster on inference. TinyBERT is also significantly better than state-of-the-art baselines on BERT di stillation, with only ~28% parameters and ~31% inference time of them.

****
UW-NET: AN INCEPTION-ATTENTION NETWORK FOR UNDERWATER IMAGE CLASSIFICATION
Miao Yang and Ke Hu,Chongyi Li,Zhiqiang Wei
The classification of images taken in special imaging environments except air is the first challenge in extending the applications of deep learning. We report o n an UW-Net (Underwater Network), a new convolutional neural network (CNN) based network for underwater image classification. In this model, we simulate the vis ual correlation of background attention with image understanding for special env ironments, such as fog and underwater by constructing an inception-attention (I-A) module. The experimental results demonstrate that the proposed UW-Net achieve s an accuracy of 99.3% on underwater image classification, which is significantl y better than other image classification networks, such as AlexNet, InceptionV3, ResNet and Se-ResNet. Moreover, we demonstrate the proposed IA module can be us ed to boost the performance of the existing object recognition networks. By subs tituting the inception module with the I-A module, the Inception-ResnetV2 networ k achieves a 10.7% top1 error rate and a 0% top5 error rate on the subset of ILS VRC-2012, which further illustrates the function of the background attention in the image classifications.
****
Semantically-Guided Representation Learning for Self-Supervised Monocular Depth
Vitor Guizilini,Rui Hou,Jie Li,Rares Ambrus,Adrien Gaidon
Self-supervised learning is showing great promise for monocular depth estimation , using geometry as the only source of supervision. Depth networks are indeed ca pable of learning representations that relate visual appearance to 3D properties by implicitly leveraging category-level patterns. In this work we investigate h ow to leverage more directly this semantic structure to guide geometric represen tation learning, while remaining in the self-supervised regime. Instead of using semantic labels and proxy losses in a multi-task approach, we propose a new arc hitecture leveraging fixed pretrained semantic segmentation networks to guide se lf-supervised representation learning via pixel-adaptive convolutions. Furthermo re, we propose a two-stage training process to overcome a common semantic bias o n dynamic objects via resampling. Our method improves upon the state of the art for self-supervised monocular depth prediction over all pixels, fine-grained det ails, and per semantic categories.

****
Stochastic AUC Maximization with Deep Neural Networks
Mingrui Liu,Zhuoning Yuan,Yiming Ying,Tianbao Yang
Stochastic AUC maximization has garnered an increasing interest due to better fi t to imbalanced data classification. However, existing works are limited to stoc hastic AUC maximization with a linear predictive model, which restricts its pred ictive power when dealing with extremely complex data. In this paper, we conside r stochastic AUC maximization problem with a deep neural network as the predicti ve model. Building on the saddle point reformulation of a surrogated loss of AUC , the problem can be cast into a {\it non-convex concave} min-max problem. The m ain contribution made in this paper is to make stochastic AUC maximization more practical for deep neural networks and big data with theoretical insights as wel l. In particular, we propose to explore Polyak-\L{}ojasiewicz (PL) condition tha t has been proved and observed in deep learning, which enables us to develop new stochastic algorithms with even faster convergence rate and more practical step size scheme. An AdaGrad-style algorithm is also analyzed under the PL condition with adaptive convergence rate. Our experimental results demonstrate the effect iveness of the proposed algorithms.
****
Data-Driven Approach to Encoding and Decoding 3-D Crystal Structures
Jordan Hoffmann,Louis Maestrati,Yoshihide Sawada,Jian Tang,Jean Michel Sellier,Y oshua Bengio
Generative models have achieved impressive results in many domains including ima ge and text generation. In the natural sciences, generative models have lead to rapid progress in automated drug discovery. Many of the current methods focus o n either 1-D or 2-D representations of typically small, drug-like molecules. How ever, many molecules require 3-D descriptors and exceed the chemical complexity of commonly used dataset. We present a method to encode and decode the position of atoms in 3-D molecules along with a dataset of nearly 50,000 stable crystal u nit cells that vary from containing 1 to over 100 atoms. We construct a smooth a nd continuous 3-D density representation of each crystal based on the positions of different atoms. Two different neural networks were trained on a dataset of o ver 120,000 three-dimensional samples of single and repeating crystal structures . The first, an Encoder-Decoder pair, constructs a compressed latent space repr esentation of each molecule and then decodes this description into an accurate r econstruction of the input. The second network segments the resulting output int o atoms and assigns each atom an atomic number. By generating compressed, contin uous latent spaces representations of molecules we are able to decode random sam ples, interpolate between two molecules, and alter known molecules.
****
Why Gradient Clipping Accelerates Training: A Theoretical Justification for Adap tivity
Jingzhao Zhang,Tianxing He,Suvrit Sra,Ali Jadbabaie
We provide a theoretical explanation for the effectiveness of gradient clipping in training deep neural networks. The key ingredient is a new smoothness conditi on derived from practical neural network training examples. We observe that grad ient smoothness, a concept central to the analysis of first-order optimization a lgorithms that is often assumed to be a constant, demonstrates significant varia bility along the training trajectory of deep neural networks. Further, this smoo thness positively correlates with the gradient norm, and contrary to standard as sumptions in the literature, it can grow with the norm of the gradient. These em pirical observations limit the applicability of existing theoretical analyses of algorithms that rely on a fixed bound on smoothness. These observations motivat e us to introduce a novel relaxation of gradient smoothness that is weaker than the commonly used Lipschitz smoothness assumption. Under the new condition, we p rove that two popular methods, namely, gradient clipping and normalized gradient , converge arbitrarily faster than gradient descent with fixed stepsize. We furt her explain why such adaptively scaled gradient methods can accelerate empirical convergence and verify our results empirically in popular neural network traini ng settings.

****
Why ADAM Beats SGD for Attention Models■
Jingzhao Zhang,Sai Praneeth Karimireddy,Andreas Veit,Seungyeon Kim,Sashank J Reddi,Sanjiv Kumar,Suvrit Sra
While stochastic gradient descent (SGD) is still the de facto algorithm in deep learning, adaptive methods like Adam have been observed to outperform SGD across important tasks, such as attention models. The settings under which SGD performs poorly in comparison to Adam are not well understood yet. In this paper, we provide empirical and theoretical evidence that a heavy-tailed distribution of the noise in stochastic gradients is a root cause of SGD's poor performance. Based on this observation, we study clipped variants of SGD that circumvent this issue; we then analyze their convergence under heavy-tailed noise. Furthermore, we develop a new adaptive coordinate-wise clipping algorithm (ACClip) tailored to such settings. Subsequently, we show how adaptive methods like Adam can be viewed through the lens of clipping, which helps us explain Adam's strong performance under heavy-tail noise settings. Finally, we show that the proposed ACClip outperforms Adam for both BERT pretraining and finetuning tasks.
****
Reflection-based Word Attribute Transfer
Yoichi Ishibashi,Katsuhito Sudoh,Koichiro Yoshino,Satoshi Nakamura
We propose a word attribute transfer framework based on reflection to obtain a word vector with an inverted target attribute for a given word in a word embedding space. Word embeddings based on Pointwise Mutual Information (PMI) represent such analogic relations as king - man + woman \approx queen. These relations can be used for changing a word’s attribute from king to queen by changing its gender. This attribute transfer can be performed by subtracting a difference vector man - woman from king when we have explicit knowledge of the gender of given word king. However, this knowledge cannot be developed for various words and attributes in practice. For transferring queen into king in this analogy-based manner, we need to know that queen denotes a female and add the difference vector to it. In this work, we transfer such binary attributes based on an assumption that such transfer mapping will become identity mapping when we apply it twice. We introduce a framework based on reflection mapping that satisfies this property; queen should be transferred back to king with the same mapping as the transfer from king to queen. Experimental results show that the proposed method can transfer the word attributes of the given words, and does not change the words that do not have the target attributes.
****
Difference-Seeking Generative Adversarial Network--Unseen Sample Generation
Yi Lin Sung,Sung-Hsien Hsieh,Soo-Chang Pei,Chun-Shien Lu

Unseen data, which are not samples from the distribution of training data and are difficult to collect, have exhibited importance in numerous applications, ({\em e.g.,} novelty detection, semi-supervised learning, and adversarial training). In this paper, we introduce a general framework called \textbf{d}ifference-\textbf{s}eeking \textbf{g}enerative \textbf{a}dversarial \textbf{n}etwork (DSGAN), to generate various types of unseen data. Its novelty is the consideration of the probability density of the unseen data distribution as the difference between two distributions $p_{\bar{d}}$ and p_{d} whose samples are relatively easy to collect.

The DSGAN can learn the target distribution, p_{t}, (or the unseen data distribution) from only the samples from the two distributions, p_{d} and $p_{\bar{d}}$. In our scenario, p_d is the distribution of the seen data, and $p_{\bar{d}}$ can be obtained from p_{d} via simple operations, so that we only need the samples of p_{d} during the training.
Two key applications, semi-supervised learning and novelty detection, are taken as case studies to illustrate that the DSGAN enables the production of various unseen data. We also provide theoretical analyses about the convergence of the DSGAN.

****
EINS: Long Short-Term Memory with Extrapolated Input Network Simplification
Nicholas I-Hsien Kuo,Mehrtash T. Harandi,Nicolas Fourrier,Gabriela Ferraro,Christian Walder,Hanna Suominen
This paper contrasts the two canonical recurrent neural networks (RNNs) of long short-term memory (LSTM) and gated recurrent unit (GRU) to propose our novel light-weight RNN of Extrapolated Input for Network Simplification (EINS). We treat LSTMs and GRUs as differential equations, and our analysis highlights several auxiliary components in the standard LSTM design that are secondary in importance. Guided by these insights, we present a design that abandons the LSTM redundancies, thereby introducing EINS. We test EINS against the LSTM over a carefully chosen range of tasks from language modelling and medical data imputation-prediction through a sentence-level variational autoencoder and image generation to learning to learn to optimise another neural network. Despite having both a simpler design and fewer parameters, this simplification either performs comparably, or better, than the LSTM in each task.
****
FasterSeg: Searching for Faster Real-time Semantic Segmentation
Wuyang Chen,Xinyu Gong,Xianming Liu,Qian Zhang,Yuan Li,Zhangyang Wang
We present FasterSeg, an automatically designed semantic segmentation network with not only state-of-the-art performance but also faster speed than current methods. Utilizing neural architecture search (NAS), FasterSeg is discovered from a novel and broader search space integrating multi-resolution branches, that has been recently found to be vital in manually designed segmentation models. To better calibrate the balance between the goals of high accuracy and low latency, we propose a decoupled and fine-grained latency regularization, that effectively overcomes our observed phenomenons that the searched networks are prone to "collapsing" to low-latency yet poor-accuracy models. Moreover, we seamlessly extend FasterSeg to a new collaborative search (co-searching) framework, simultaneously searching for a teacher and a student network in the same single run. The teacher-student distillation further boosts the student model’s accuracy. Experiments on popular segmentation benchmarks demonstrate the competency of FasterSeg. For example, FasterSeg can run over 30% faster than the closest manually designed competitor on Cityscapes, while maintaining comparable accuracy.
****
LEARNING EXECUTION THROUGH NEURAL CODE FUSION
Zhan Shi,Kevin Swersky,Daniel Tarlow,Parthasarathy Ranganathan,Milad Hashemi
As the performance of computer systems stagnates due to the end of Moore’s Law, there is a need for new models that can understand and optimize the execution of general purpose code. While there is a growing body of work on using Graph Neural Networks (GNNs) to learn static representations of source code, these representations do not understand how code executes at runtime. In this work, we propose a new approach using GNNs to learn fused representations of general source code and its execution. Our approach defines a multi-task GNN over low-level representations of source code and program state (i.e., assembly code and dynamic memory states), converting complex source code constructs and data structures into a simpler, more uniform format. We show that this leads to improved
performance over similar methods that do not use execution and it opens the door to applying GNN models to new tasks that would not be feasible from static code alone. As an illustration of this, we apply the new model to challenging dynamic tasks (branch prediction and prefetching) from the SPEC CPU benchmark suite, outperforming the state-of-the-art by 26% and 45% respectively. Moreover, we use the learned fused graph embeddings to demonstrate transfer learning with high
performance on an indirectly related algorithm classification task.
****
Editable Neural Networks

Anton Sinitsin,Vsevolod Plokhotnyuk,Dmitry Pyrkin,Sergei Popov,Artem Babenko
These days deep neural networks are ubiquitously used in a wide range of tasks, from image classification and machine translation to face identification and self-driving cars. In many applications, a single model error can lead to devastating financial, reputational and even life-threatening consequences. Therefore, it is crucially important to correct model mistakes quickly as they appear. In this work, we investigate the problem of neural network editing - how one can efficiently patch a mistake of the model on a particular sample, without influencing the model behavior on other samples. Namely, we propose Editable Training, a model-agnostic training technique that encourages fast editing of the trained model. We empirically demonstrate the effectiveness of this method on large-scale image classification and machine translation tasks.
****

Parallel Scheduled Sampling
Daniel Duckworth,Arvind Neelakantan,Ben Goodrich,Lukasz Kaiser,Samy Bengio
Auto-regressive models are widely used in sequence generation problems. The output sequence is typically generated in a predetermined order, one discrete unit(pixel or word or character) at a time. The models are trained by teacher-forcing where ground-truth history is fed to the model as input, which at test time is replaced by the model prediction. Scheduled Sampling (Bengio et al., 2015) aims to mitigate this discrepancy between train and test time by randomly replacing some discrete units in the history with the model’s prediction. While teacher-forced training works well with ML accelerators as the computation can be parallelized across time, Scheduled Sampling involves undesirable sequential processing. In this paper, we introduce a simple technique to parallelize Scheduled Sampling across time. Experimentally, we find the proposed technique leads to equivalent or better performance on image generation, summarization, dialog generation, and translation compared to teacher-forced training. n dialog response generation task,Parallel Scheduled Sampling achieves 1.6 BLEU score (11.5%) improvement over teacher-forcing while in image generation it achieves 20% and 13.8% improvement in Frechet Inception Distance (FID) and Inception Score (IS) respectively. Further, we discuss the effects of different hyper-parameters associated with Scheduled Sampling on the model performance.
****

Learning Explainable Models Using Attribution Priors
Gabriel Erion,Joseph D. Janizek,Pascal Sturmfels,Scott M. Lundberg,Su-In Lee
Two important topics in deep learning both involve incorporating humans into the modeling process: Model priors transfer information from humans to a model by regularizing the model's parameters; Model attributions transfer information from a model to humans by explaining the model's behavior. Previous work has taken important steps to connect these topics through various forms of gradient regularization. We find, however, that existing methods that use attributions to align a model's behavior with human intuition are ineffective. We develop an efficient and theoretically grounded feature attribution method, expected gradients, and a novel framework, attribution priors, to enforce prior expectations about a model's behavior during training. We demonstrate that attribution priors are broadly applicable by instantiating them on three different types of data: image data, gene expression data, and health care data. Our experiments show that models trained with attribution priors are more intuitive and achieve better generalization performance than both equivalent baselines and existing methods to regularize model behavior.
****

Efficient Inference and Exploration for Reinforcement Learning
Yi Zhu,Jing Dong,Henry Lam
Despite an ever growing literature on reinforcement learning algorithms and applications, much less is known about their statistical inference. In this paper, we investigate the large-sample behaviors of the Q-value estimates with closed-form characterizations of the asymptotic variances. This allows us to efficiently construct confidence regions for Q-value and optimal value functions, and to develop policies to minimize their estimation errors. This also leads to a policy e

xploration strategy that relies on estimating the relative discrepancies among t
he Q estimates. Numerical experiments show superior performances of our explorat
ion strategy than other benchmark approaches.
****

Leveraging inductive bias of neural networks for learning without explicit human
 annotations
Fatih Furkan Yilmaz,Reinhard Heckel
Classification problems today are typically solved by first collecting examples
along with candidate labels, second obtaining clean labels from workers,
and third training a large, overparameterized deep neural network on the clean e
xamples. The second, labeling step is often the most expensive one as it require
s manually going through all examples.
In this paper we skip the labeling step entirely and propose to directly train t
he deep neural network on the noisy raw labels and early stop the training to av
oid overfitting.
With this procedure we exploit an intriguing property of large overparameterized
 neural networks: While they are capable of perfectly fitting the noisy data, gr
adient descent fits clean labels much faster than the noisy ones, thus early sto
pping resembles training on the clean labels.
Our results show that early stopping the training of standard deep networks such
 as ResNet-18 on part of the Tiny Images dataset, which does not involve any hum
an labeled data, and of which only about half of the labels are correct, gives a
 significantly higher test performance than when trained on the clean CIFAR-10 t
raining dataset, which is a labeled version of the Tiny Images dataset, for the
same classification problem.
In addition, our results show that the noise generated through the label collect
ion process is not nearly as adversarial for learning as the noise generated by
randomly flipping labels, which is the noise most prevalent in works demonstrati
ng noise robustness of neural networks.
****

Bias-Resilient Neural Network
Ehsan Adeli,Qingyu Zhao,Adolf Pfefferbaum,Edith V. Sullivan,L. Fei-Fei,Juan Carl
os Niebles,Kilian M. Pohl
Presence of bias and confounding effects is inarguably one of the most critical
challenges in machine learning applications that has alluded to pivotal debates
in the recent years. Such challenges range from spurious associations of confoun
ding variables in medical studies to the bias of race in gender or face recognit
ion systems. One solution is to enhance datasets and organize them such that the
y do not reflect biases, which is a cumbersome and intensive task. The alternati
ve is to make use of available data and build models considering these biases. T
raditional statistical methods apply straightforward techniques such as residual
ization or stratification to precomputed features to account for confounding var
iables. However, these techniques are not in general applicable to end-to-end de
ep learning methods. In this paper, we propose a method based on the adversarial
 training strategy to learn discriminative features unbiased and invariant to th
e confounder(s). This is enabled by incorporating a new adversarial loss functio
n that encourages a vanished correlation between the bias and learned features.
We apply our method to a synthetic, a medical diagnosis, and a gender classifica
tion (Gender Shades) dataset. Our results show that the learned features by our
method not only result in superior prediction performance but also are uncorrela
ted with the bias or confounder variables. The code is available at http://blind
ed_for_review/.
****

Effective Use of Variational Embedding Capacity in Expressive End-to-End Speech
Synthesis
Eric Battenberg,Soroosh Mariooryad,Daisy Stanton,RJ Skerry-Ryan,Matt Shannon,Dav
id Kao,Tom Bagby
Recent work has explored sequence-to-sequence latent variable models for express
ive speech synthesis (supporting control and transfer of prosody and style), but
 has not presented a coherent framework for understanding the trade-offs between

the competing methods. In this paper, we propose embedding capacity (the amount of information the embedding contains about the data) as a unified method of analyzing the behavior of latent variable models of speech, comparing existing heuristic (non-variational) methods to variational methods that are able to explicitly constrain capacity using an upper bound on representational mutual information. In our proposed model (Capacitron), we show that by adding conditional dependencies to the variational posterior such that it matches the form of the true posterior, the same model can be used for high-precision prosody transfer, text-agnostic style transfer, and generation of natural-sounding prior samples. For multi-speaker models, Capacitron is able to preserve target speaker identity during inter-speaker prosody transfer and when drawing samples from the latent prior. Lastly, we introduce a method for decomposing embedding capacity hierarchically across two sets of latents, allowing a portion of the latent variability to be specified and the remaining variability sampled from a learned prior. Audio examples are available on the web.

****

Accelerating Reinforcement Learning Through GPU Atari Emulation

Steven Dalton,Michael Garland,Iuri Frosio

We introduce CuLE (CUDA Learning Environment), a CUDA port of the Atari Learning Environment (ALE) which is used for the development of deep reinforcement algorithms. CuLE overcomes many limitations of existing CPU-based emulators and scales naturally to multiple GPUs. It leverages GPU parallelization to run thousands of games simultaneously and it renders frames directly on the GPU, to avoid the bottleneck arising from the limited CPU-GPU communication bandwidth. CuLE generates up to 155M frames per hour on a single GPU, a finding previously achieved only through a cluster of CPUs. Beyond highlighting the differences between CPU and GPU emulators in the context of reinforcement learning, we show how to leverage the high throughput of CuLE by effective batching of the training data, and show accelerated convergence for A2C+V-trace. CuLE is available at [hidden URL].

****

Can gradient clipping mitigate label noise?

Aditya Krishna Menon,Ankit Singh Rawat,Sashank J. Reddi,Sanjiv Kumar

Gradient clipping is a widely-used technique in the training of deep networks, and is generally motivated from an optimisation lens: informally, it controls the dynamics of iterates, thus enhancing the rate of convergence to a local minimum. This intuition has been made precise in a line of recent works, which show that suitable clipping can yield significantly faster convergence than vanilla gradient descent. In this paper, we propose a new lens for studying gradient clipping, namely, robustness: informally, one expects clipping to provide robustness to noise, since one does not overly trust any single sample. Surprisingly, we prove that for the common problem of label noise in classification, standard gradient clipping does not in general provide robustness. On the other hand, we show that a simple variant of gradient clipping is provably robust, and corresponds to suitably modifying the underlying loss function. This yields a simple, noise-robust alternative to the standard cross-entropy loss which performs well empirically.

****

Concise Multi-head Attention Models

Srinadh Bhojanapalli,Chulhee Yun,Ankit Singh Rawat,Sashank Reddi,Sanjiv Kumar

Attention based Transformer architecture has enabled significant advances in the field of natural language processing. In addition to new pre-training techniques, recent improvements crucially rely on working with a relatively larger embedding dimension for tokens. This leads to models that are prohibitively large to be employed in the downstream tasks. In this paper we identify one of the important factors contributing to the large embedding size requirement. In particular, our analysis highlights that the scaling between the number of heads and the size of each head in the existing architectures gives rise to this limitation, which we further validate with our experiments. As a solution, we propose a new way to set the projection size in attention heads that allows us to train models with a relatively smaller embedding dimension, without sacrificing the performance

.
****
Tensorized Embedding Layers for Efficient Model Compression
Oleksii Hrinchuk,Valentin Khrulkov,Leyla Mirvakhabova,Ivan Oseledets
The embedding layers transforming input words into real vectors are the key components of deep neural networks used in natural language processing. However, when the vocabulary is large, the corresponding weight matrices can be enormous, which precludes their deployment in a limited resource setting. We introduce a novel way of parametrizing embedding layers based on the Tensor Train (TT) decomposition, which allows compressing the model significantly at the cost of a negligible drop or even a slight gain in performance. We evaluate our method on a wide range of benchmarks in natural language processing and analyze the trade-off between performance and compression ratios for a wide range of architectures, from MLPs to LSTMs and Transformers.
****
Rethinking Neural Network Quantization
Qing Jin,Linjie Yang,Zhenyu Liao
Quantization reduces computation costs of neural networks but suffers from performance degeneration. Is this accuracy drop due to the reduced capacity, or inefficient training during the quantization procedure? After looking into the gradient propagation process of neural networks by viewing the weights and intermediate activations as random variables, we discover two critical rules for efficient training. Recent quantization approaches violates the two rules and results in degenerated convergence. To deal with this problem, we propose a simple yet effective technique, named scale-adjusted training (SAT), to comply with the discovered rules and facilitates efficient training. We also analyze the quantization error introduced in calculating the gradient in the popular parameterized clipping activation (PACT) technique. Through SAT together with gradient-calibrated PACT, quantized models obtain comparable or even better performance than their full-precision counterparts, achieving state-of-the-art accuracy with consistent improvement over previous quantization methods on a wide spectrum of models including MobileNet-V1/V2 and PreResNet-50.
****
Zero-shot task adaptation by homoiconic meta-mapping
Andrew K. Lampinen,James L. McClelland
How can deep learning systems flexibly reuse their knowledge? Toward this goal, we propose a new class of challenges, and a class of architectures that can solve them. The challenges are meta-mappings, which involve systematically transforming task behaviors to adapt to new tasks zero-shot. We suggest that the key to achieving these challenges is representing the task being performed in such a way that this task representation is itself transformable. We therefore draw inspiration from functional programming and recent work in meta-learning to propose a class of Homoiconic Meta-Mapping (HoMM) approaches that represent data points and tasks in a shared latent space, and learn to infer transformations of that space. HoMM approaches can be applied to any type of machine learning task, including supervised learning and reinforcement learning. We demonstrate the utility of this perspective by exhibiting zero-shot remapping of behavior to adapt to new tasks.
****
iSparse: Output Informed Sparsification of Neural Networks
Yash Garg,K. Selcuk Candan
Deep neural networks have demonstrated unprecedented success in various knowledge management applications. However, the networks created are often very complex, with large numbers of trainable edges which require extensive computational resources. We note that many successful networks nevertheless often contain large numbers of redundant edges. Moreover, many of these edges may have negligible contributions towards the overall network performance. In this paper, we propose a novel iSparse framework and experimentally show, that we can sparsify the network, by 30-50%, without impacting the network performance. iSparse leverages a novel edge significance score, E, to determine the importance of an edge with respe

ct to the final network output. Furthermore, iSparse can be applied both while t raining a model or on top of a pre-trained model, making it a retraining-free a pproach - leading to a minimal computational overhead. Comparisons of iSparse ag ainst PFEC, NISP, DropConnect, and Retraining-Free on benchmark datasets show th at iSparse leads to effective network sparsifications.
****

HyperEmbed: Tradeoffs Between Resources and Performance in NLP Tasks with Hyper dimensional Computing enabled embedding of n-gram statistics
Pedro Alonso,Kumar Shridhar,Denis Kleyko,Evgeny Osipov,Marcus Liwicki
Recent advances in Deep Learning have led to a significant performance increase on several NLP tasks, however, the models become more and more computationally d emanding. Therefore, this paper tackles the domain of computationally efficient algorithms for NLP tasks. In particular, it investigates distributed representat ions of n-gram statistics of texts. The representations are formed using hyperdi mensional computing enabled embedding. These representations then serve as featu res, which are used as input to standard classifiers. We investigate the applica bility of the embedding on one large and three small standard datasets for class ification tasks using nine classifiers. The embedding achieved on par F1 scores while decreasing the time and memory requirements by several times compared to the conventional n-gram statistics, e.g., for one of the classifiers on a small dataset, the memory reduction was 6.18 times; while train and test speed-ups wer e 4.62 and 3.84 times, respectively. For many classifiers on the large dataset, the memory reduction was about 100 times and train and test speed-ups were over 100 times. More importantly, the usage of distributed representations formed via hyperdimensional computing allows dissecting the strict dependency between the dimensionality of the representation and the parameters of n-gram statistics, th us, opening a room for tradeoffs.
****

Pretrained Encyclopedia: Weakly Supervised Knowledge-Pretrained Language Model
Wenhan Xiong,Jingfei Du,William Yang Wang,Veselin Stoyanov
 Recent breakthroughs of pretrained language models have shown the effectiveness of self-supervised learning for a wide range of natural language processing (NL P) tasks. In addition to standard syntactic and semantic NLP tasks, pretrained m odels achieve strong improvements on tasks that involve real-world knowledge, su ggesting that large-scale language modeling could be an implicit method to captu re knowledge. In this work, we further investigate the extent to which pretraine d models such as BERT capture knowledge using a zero-shot fact completion task. Moreover, we propose a simple yet effective weakly supervised pretraining object ive, which explicitly forces the model to incorporate knowledge about real-world entities. Models trained with our new objective yield significant improvements on the fact completion task. When applied to downstream tasks, our model consist ently outperforms BERT on four entity-related question answering datasets (i.e., WebQuestions, TriviaQA, SearchQA and Quasar-T) with an average 2.7 F1 improveme nts and a standard fine-grained entity typing dataset (i.e., FIGER) with 5.7 acc uracy gains.
****

Fast Linear Interpolation for Piecewise-Linear Functions, GAMs, and Deep Lattice Networks
Nathan Zhang,Kevin Canini,Sean Silva,and Maya R. Gupta
We present fast implementations of linear interpolation operators for both piece wise linear functions and multi-dimensional look-up tables. We use a compiler-ba sed solution (using MLIR) for accelerating this family of workloads. On real-wor ld multi-layer lattice models and a standard CPU, we show these strategies deliv er $5-10\times$ faster runtimes compared to a C++ interpreter implementation tha t uses prior techniques, producing runtimes that are 1000s of times faster than TensorFlow 2.0 for single evaluations.
****

Adversarial Training: embedding adversarial perturbations into the parameter spa ce of a neural network to build a robust system
Shixian Wen,Laurent Itti

Adversarial training, in which a network is trained on both adversarial and clea
n examples, is one of the most trusted defense methods against adversarial attac
ks. However, there are three major practical difficulties in implementing and de
ploying this method - expensive in terms of extra memory and computation costs;
accuracy trade-off between clean and adversarial examples; and lack of diversity
 of adversarial perturbations. Classical adversarial training uses fixed, precom
puted perturbations in adversarial examples (input space). In contrast, we intro
duce dynamic adversarial perturbations into the parameter space of the network,
by adding perturbation biases to the fully connected layers of deep convolutiona
l neural network. During training, using only clean images, the perturbation bia
ses are updated in the Fast Gradient Sign Direction to automatically create and
store adversarial perturbations by recycling the gradient information computed.
The network learns and adjusts itself automatically to these learned adversarial
 perturbations. Thus, we can achieve adversarial training with negligible cost c
ompared to requiring a training set of adversarial example images. In addition,
if combined with classical adversarial training, our perturbation biases can all
eviate accuracy trade-off difficulties, and diversify adversarial perturbations.
****

Collaborative Generated Hashing for Market Analysis and Fast Cold-start Recommen
dation

Yan Zhang,Ivor W. Tsang,Lixin Duan,Guowu Yang

Cold-start and efficiency issues of the Top-k recommendation are critical to lar
ge-scale recommender systems. Previous hybrid recommendation methods are effecti
ve to deal with the cold-start issues by extracting real latent factors of cold-
start items(users) from side information, but they still suffer low efficiency i
n online recommendation caused by the expensive similarity search in real latent
 space. This paper presents a collaborative generated hashing (CGH) to improve t
he efficiency by denoting users and items as binary codes, which applies to vari
ous settings: cold-start users, cold-start items and warm-start ones. Specifical
ly, CGH is designed to learn hash functions of users and items through the Minim
um Description Length (MDL) principle; thus, it can deal with various recommenda
tion settings. In addition, CGH initiates a new marketing strategy through minin
g potential users by a generative step. To reconstruct effective users, the MDL
principle is used to learn compact and informative binary codes from the content
 data. Extensive experiments on two public datasets show the advantages for reco
mmendations in various settings over competing baselines and analyze the feasibi
lity of the application in marketing.
****

Pruned Graph Scattering Transforms

Vassilis N. Ioannidis,Siheng Chen,Georgios B. Giannakis

Graph convolutional networks (GCNs) have achieved remarkable performance in a va
riety of network science learning tasks. However, theoretical analysis of such a
pproaches is still at its infancy. Graph scattering transforms (GSTs) are non-tr
ainable deep GCN models that are amenable to generalization and stability analys
es. The present work addresses some limitations of GSTs by introducing a novel s
o-termed pruned (p)GST approach. The resultant pruning algorithm is guided by a
graph-spectrum-inspired criterion, and retains informative scattering features o
n-the-fly while bypassing the exponential complexity associated with GSTs. It is
 further established that pGSTs are stable to perturbations of the input graph s
ignals with bounded energy. Experiments showcase that i) pGST performs comparabl
y to the baseline GST that uses all scattering features, while achieving signifi
cant computational savings; ii) pGST achieves comparable performance to state-of
-the-art GCNs; and iii) Graph data from various domains lead to different scatte
ring patterns, suggesting domain-adaptive pGST network architectures.
****

DDSP: Differentiable Digital Signal Processing

Jesse Engel,Lamtharn (Hanoi) Hantrakul,Chenjie Gu,Adam Roberts

Most generative models of audio directly generate samples in one of two domains:
 time or frequency. While sufficient to express any signal, these representation
s are inefficient, as they do not utilize existing knowledge of how sound is gen

erated and perceived. A third approach (vocoders/synthesizers) successfully incorporates strong domain knowledge of signal processing and perception, but has been less actively researched due to limited expressivity and difficulty integrating with modern auto-differentiation-based machine learning methods. In this paper, we introduce the Differentiable Digital Signal Processing (DDSP) library, which enables direct integration of classic signal processing elements with deep learning methods. Focusing on audio synthesis, we achieve high-fidelity generation without the need for large autoregressive models or adversarial losses, demonstrating that DDSP enables utilizing strong inductive biases without losing the expressive power of neural networks. Further, we show that combining interpretable modules permits manipulation of each separate model component, with applications such as independent control of pitch and loudness, realistic extrapolation to pitches not seen during training, blind dereverberation of room acoustics, transfer of extracted room acoustics to new environments, and transformation of timbre between disparate sources. In short, DDSP enables an interpretable and modular approach to generative modeling, without sacrificing the benefits of deep learning. The library will is available at https://github.com/magenta/ddsp and we encourage further contributions from the community and domain experts.

****

Continual Learning via Neural Pruning

Siavash Golkar,Micheal Kagan,Kyunghyun Cho

We introduce Continual Learning via Neural Pruning~(CLNP), a new method aimed at lifelong learning in fixed capacity models based on neuronal model sparsification. In this method, subsequent tasks are trained using the inactive neurons and filters of the sparsified network and cause zero deterioration to the performance of previous tasks. In order to deal with the possible compromise between model sparsity and performance, we formalize and incorporate the concept of \emph{graceful forgetting}: the idea that it is preferable to suffer a small amount of forgetting in a controlled manner if it helps regain network capacity and prevents uncontrolled loss of performance during the training of future tasks. CLNP also provides simple continual learning diagnostic tools in terms of the number of free neurons left for the training of future tasks as well as the number of neurons that are being reused. In particular, we see in experiments that CLNP verifies and automatically takes advantage of the fact that the features of earlier layers are more transferable. We show empirically that CLNP leads to significantly improved results over current weight elasticity based methods. CLNP can also be applied in single-head architectures providing the first viable such algorithm for continual learning.

****

Min-Max Optimization without Gradients: Convergence and Applications to Adversarial ML

Sijia Liu,Songtao Lu,Xiangyi Chen,Yao Feng,Kaidi Xu,Abdullah Al-Dujaili,Minyi Hong,Una-May Obelilly

In this paper, we study the problem of constrained robust (min-max) optimization ina black-box setting, where the desired optimizer cannot access the gradients of the objective function but may query its values. We present a principled optimization framework, integrating a zeroth-order (ZO) gradient estimator with an alternating projected stochastic gradient descent-ascent method, where the former only requires a small number of function queries and the later needs just one-step descent/ascent update. We show that the proposed framework, referred to as ZO-Min-Max, has a sub-linear convergence rate under mild conditions and scales gracefully with problem size. From an application side, we explore a promising connection between black-box min-max optimization and black-box evasion and poisoning attacks in adversarial machine learning (ML). Our empirical evaluations on these use cases demonstrate the effectiveness of our approach and its scalability to dimensions that prohibit using recent black-box solvers.

****

XLDA: Cross-Lingual Data Augmentation for Natural Language Inference and Question Answering

Jasdeep Singh,Bryan McCann,Nitish Shirish Keskar,Caiming Xiong,Richard Socher
While natural language processing systems often focus on a single language, mult
ilingual transfer learning has the potential to improve performance, especially
for low-resource languages.
We introduce XLDA, cross-lingual data augmentation, a method that replaces a seg
ment of the input text with its translation in another language. XLDA enhances p
erformance of all 14 tested languages of the cross-lingual natural language infe
rence (XNLI) benchmark. With improvements of up to 4.8, training with XLDA achie
ves state-of-the-art performance for Greek, Turkish, and Urdu. XLDA is in contra
st to, and performs markedly better than, a more naive approach that aggregates
examples in various languages in a way that each example is solely in one langua
ge. On the SQuAD question answering task, we see that XLDA provides a 1.0 perfor
mance increase on the English evaluation set. Comprehensive experiments suggest
that most languages are effective as cross-lingual augmentors, that XLDA is robu
st to a wide range of translation quality, and that XLDA is even more effective
for randomly initialized models than for pretrained models.
****

Attraction-Repulsion Actor-Critic for Continuous Control Reinforcement Learning
Thang Doan,Bogdan Mazoure,Audrey Durand,Joelle Pineau,R Devon Hjelm
Continuous control tasks in reinforcement learning are important because they pr
ovide an important framework for learning in high-dimensional state spaces with
deceptive rewards, where the agent can easily become trapped into suboptimal sol
utions.
One way to avoid local optima is to use a population of agents to ensure coverag
e of the policy space, yet learning a population with the ``best" coverage is st
ill an open problem. In this work, we present a novel approach to population-bas
ed RL in continuous control that leverages properties of normalizing flows to pe
rform attractive and repulsive operations between current members of the populat
ion and previously observed policies. Empirical results on the MuJoCo suite demo
nstrate a high performance gain for our algorithm compared to prior work, includ
ing Soft-Actor Critic (SAC).
****

GLAD: Learning Sparse Graph Recovery
Harsh Shrivastava,Xinshi Chen,Binghong Chen,Guanghui Lan,Srinivas Aluru,Han Liu,
Le Song
Recovering sparse conditional independence graphs from data is a fundamental pro
blem in machine learning with wide applications. A popular formulation of the pr
oblem is an ℓ_1 regularized maximum likelihood estimation. Many convex opti
mization algorithms have been designed to solve this formulation to recover the
graph structure. Recently, there is a surge of interest to learn algorithms dire
ctly based on data, and in this case, learn to map empirical covariance to the s
parse precision matrix. However, it is a challenging task in this case, since th
e symmetric positive definiteness (SPD) and sparsity of the matrix are not easy
to enforce in learned algorithms, and a direct mapping from data to precision ma
trix may contain many parameters. We propose a deep learning architecture, GLAD,
 which uses an Alternating Minimization (AM) algorithm as our model inductive bi
as, and learns the model parameters via supervised learning. We show that GLAD l
earns a very compact and effective model for recovering sparse graphs from data.
****

PDP: A General Neural Framework for Learning SAT Solvers
Saeed Amizadeh,Sergiy Matusevych,Markus Weimer
There have been recent efforts for incorporating Graph Neural Network models for
 learning fully neural solvers for constraint satisfaction problems (CSP) and pa
rticularly Boolean satisfiability (SAT). Despite the unique representational pow
er of these neural embedding models, it is not clear to what extent they actuall
y learn a search strategy vs. statistical biases in the training data. On the ot
her hand, by fixing the search strategy (e.g. greedy search), one would effectiv
ely deprive the neural models of learning better strategies than those given. In
 this paper, we propose a generic neural framework for learning SAT solvers (and
 in general any CSP solver) that can be described in terms of probabilistic infe

rence and yet learn search strategies beyond greedy search. Our framework is bas ed on the idea of propagation, decimation and prediction (and hence the name PDP) in graphical models, and can be trained directly toward solving SAT in a fully unsupervised manner via energy minimization, as shown in the paper. Our experim ental results demonstrate the effectiveness of our framework for SAT solving com pared to both neural and the industrial baselines.
****

Adaptive Loss Scaling for Mixed Precision Training
Ruizhe Zhao,Brian Vogel,Tanvir Ahmed
Mixed precision training (MPT) is becoming a practical technique to improve the speed and energy efficiency of training deep neural networks by leveraging the f ast hardware support for IEEE half-precision floating point that is available in existing GPUs. MPT is typically used in combination with a technique called los s scaling, that works by scaling up the loss value up before the start of backpr opagation in order to minimize the impact of numerical underflow on training. Un fortunately, existing methods make this loss scale value a hyperparameter that n eeds to be tuned per-model, and a single scale cannot be adapted to different la yers at different training stages. We introduce a loss scaling-based training me thod called adaptive loss scaling that makes MPT easier and more practical to us e, by removing the need to tune a model-specific loss scale hyperparameter. We a chieve this by introducing layer-wise loss scale values which are automatically computed during training to deal with underflow more effectively than existing m ethods. We present experimental results on a variety of networks and tasks that show our approach can shorten the time to convergence and improve accuracy, comp ared with using the existing state-of-the-art MPT and single-precision floating point.
****

How many weights are enough : can tensor factorization learn efficient policies ?
Pierre H. Richemond,Arinbjorn Kolbeinsson,Yike Guo
Deep reinforcement learning requires a heavy price in terms of sample efficiency and overparameterization in the neural networks used for function approximation . In this work, we employ tensor factorization in order to learn more compact re presentations for reinforcement learning policies. We show empirically that in t he low-data regime, it is possible to learn online policies with 2 to 10 times l ess total coefficients, with little to no loss of performance. We also leverage progress in second order optimization, and use the theory of wavelet scattering to further reduce the number of learned coefficients, by foregoing learning the topmost convolutional layer filters altogether. We evaluate our results on the A tari suite against recent baseline algorithms that represent the state-of-the-ar t in data efficiency, and get comparable results with an order of magnitude gain in weight parsimony.
****

Domain Aggregation Networks for Multi-Source Domain Adaptation
Junfeng Wen,Russell Greiner,Dale Schuurmans
In many real-world applications, we want to exploit multiple source datasets of similar tasks to learn a model for a different but related target dataset -- e.g ., recognizing characters of a new font using a set of different fonts. While m ost recent research has considered ad-hoc combination rules to address this prob lem, we extend previous work on domain discrepancy minimization to develop a fin ite-sample generalization bound, and accordingly propose a theoretically justifi ed optimization procedure. The algorithm we develop, Domain AggRegation Network (DARN), is able to effectively adjust the weight of each source domain during tr aining to ensure relevant domains are given more importance for adaptation. We e valuate the proposed method on real-world sentiment analysis and digit recogniti on datasets and show that DARN can significantly outperform the state-of-the-art alternatives.
****

Benchmarking Robustness in Object Detection: Autonomous Driving when Winter is C oming

Claudio Michaelis,Benjamin Mitzkus,Robert Geirhos,Evgenia Rusak,Oliver Bringmann ,Alexander S. Ecker,Matthias Bethge,Wieland Brendel
The ability to detect objects regardless of image distortions or weather conditi ons is crucial for real-world applications of deep learning like autonomous driv ing. We here provide an easy-to-use benchmark to assess how object detection mod els perform when image quality degrades. The three resulting benchmark datasets, termed PASCAL-C, COCO-C and Cityscapes-C, contain a large variety of image corr uptions. We show that a range of standard object detection models suffer a sever e performance loss on corrupted images (down to 30-60% of the original performan ce). However, a simple data augmentation trick - stylizing the training images - leads to a substantial increase in robustness across corruption type, severity and dataset. We envision our comprehensive benchmark to track future progress to wards building robust object detection models. Benchmark, code and data are avai lable at: (hidden for double blind review)
****

AHash: A Load-Balanced One Permutation Hash
Chenxingyu Zhao,Jie Gui,Yixiao Guo,Jie Jiang,Tong Yang,Bin Cui,Gong Zhang
Minwise Hashing (MinHash) is a fundamental method to compute set similarities an d compact high-dimensional data for efficient learning and searching. The bottle neck of MinHash is computing k (usually hundreds) MinHash values. One Permutatio n Hashing (OPH) only requires one permutation (hash function) to get k MinHash values by dividing elements into k bins. One drawback of OPH is that the load o f the bins (the number of elements in a bin) could be unbalanced, which leads to the existence of empty bins and false similarity computation. Several strategie s for densification, that is, filling empty bins, have been proposed. However, t he densification is just a remedial strategy and cannot eliminate the error incu rred by the unbalanced load. Unlike the densification to fill the empty bins aft er they undesirably occur, our design goal is to balance the load so as to reduc e the empty bins in advance. In this paper, we propose a load-balanced hashing, Amortization Hashing (AHash), which can generate as few empty bins as possible. Therefore, AHash is more load-balanced and accurate without hurting runtime eff iciency compared with OPH and densification strategies. Our experiments on real datasets validate the claim. All source codes and datasets have been provided as Supplementary Materials and released on GitHub anonymously.
****

Ordinary differential equations on graph networks
Juntang Zhuang,Nicha Dvornek,Xiaoxiao Li,James S. Duncan
Recently various neural networks have been proposed for irregularly structured d ata such as graphs and manifolds. To our knowledge, all existing graph networks have discrete depth. Inspired by neural ordinary differential equation (NODE) fo r data in the Euclidean domain, we extend the idea of continuous-depth models to graph data, and propose graph ordinary differential equation (GODE). The deriva tive of hidden node states are parameterized with a graph neural network, and th e output states are the solution to this ordinary differential equation. We demo nstrate two end-to-end methods for efficient training of GODE: (1) indirect back -propagation with the adjoint method; (2) direct back-propagation through the OD E solver, which accurately computes the gradient. We demonstrate that direct bac kprop outperforms the adjoint method in experiments. We then introduce a family of bijective blocks, which enables $\mathcal{O}(1)$ memory consumption. We demon strate that GODE can be easily adapted to different existing graph neural networ ks and improve accuracy. We validate the performance of GODE in both semi-superv ised node classification tasks and graph classification tasks. Our GODE model ac hieves a continuous model in time, memory efficiency, accurate gradient estimati on, and generalizability with different graph networks.
****

Lift-the-flap: what, where and when for context reasoning
Mengmi Zhang,Claire Tseng,Karla Montejo,Joseph Kwon,Gabriel Kreiman
Context reasoning is critical in a wide variety of applications where current in puts need to be interpreted in the light of previous experience and knowledge. B oth spatial and temporal contextual information play a critical role in the doma

in of visual recognition. Here we investigate spatial constraints (what image fe atures provide contextual information and where they are located), and temporal constraints (when different contextual cues matter) for visual recognition. The task is to reason about the scene context and infer what a target object hidden behind a flap is in a natural image. To tackle this problem, we first describe a n online human psychophysics experiment recording active sampling via mouse clic ks in lift-the-flap games and identify clicking patterns and features which are diagnostic for high contextual reasoning accuracy. As a proof of the usefulness of these clicking patterns and visual features, we extend a state-of-the-art rec urrent model capable of attending to salient context regions, dynamically integr ating useful information, making inferences, and predicting class label for the target object over multiple clicks. The proposed model achieves human-level cont extual reasoning accuracy, shares human-like sampling behavior and learns interp retable features for contextual reasoning.
****

Unifying Question Answering, Text Classification, and Regression via Span Extrac tion
Nitish Shirish Keskar,Bryan McCann,Caiming Xiong,Richard Socher
Even as pre-trained language encoders such as BERT are shared across many tasks, the output layers of question answering, text classification, and regression mo dels are significantly different. Span decoders are frequently used for question answering, fixed-class, classification layers for text classification, and simi larity-scoring layers for regression tasks. We show that this distinction is not necessary and that all three can be unified as span extraction. A unified, span -extraction approach leads to superior or comparable performance in supplementar y supervised pre-trained, low-data, and multi-task learning experiments on seve ral question answering, text classification, and regression benchmarks.
****

Supervised learning with incomplete data via sparse representations
Cesar F. Caiafa,Ziyao Wang,Jordi Solé-Casals,Qibin Zhao
This paper addresses the problem of training a classifier on incomplete data and its application to a complete or incomplete test dataset. A supervised learning method is developed to train a general classifier, such as a logistic regressio n or a deep neural network, using only a limited number of observed entries, ass uming sparse representations of data vectors on an unknown dictionary. The propo sed method simultaneously learns the classifier, the dictionary and the correspo nding sparse representations of each input data sample. A theoretical analysis i s also provided comparing this method with the standard imputation approach, whi ch consists on performing data completion followed by training the classifier ba sed on their reconstructions. The limitations of this last "sequential" approach are identified, and a description of how the proposed new "simultaneous" method can overcome the problem of indiscernible observations is provided. Additionall y, it is shown that, if it is possible to train a classifier on incomplete obser vations so that its reconstructions are well separated by a hyperplane, then the same classifier also correctly separates the original (unobserved) data samples . Extensive simulation results are presented on synthetic and well-known referen ce datasets that demonstrate the effectiveness of the proposed method compared t o traditional data imputation methods.
****

The Probabilistic Fault Tolerance of Neural Networks in the Continuous Limit
El-Mahdi El-Mhamdi,Rachid Guerraoui,Andrei Kucharavy,Sergei Volodin
The loss of a few neurons in a brain rarely results in any visible loss of funct ion. However, the insight into what “few” means in this context is unclear. How many random neuron failures will it take to lead to a visible loss of function? In this paper, we address the fundamental question of the impact of the crash of a random subset of neurons on the overall computation of a neural network and t he error in the output it produces. We study fault tolerance of neural networks subject to small random neuron/weight crash failures in a probabilistic setting. We give provable guarantees on the robustness of the network to these crashes. Our main contribution is a bound on the error in the output of a network under s

mall random Bernoulli crashes proved by using a Taylor expansion in the continuo us limit, where close-by neurons at a layer are similar. The failure mode we ado pt in our model is characteristic of neuromorphic hardware, a promising technolo gy to speed up artificial neural networks, as well as of biological networks. We show that our theoretical bounds can be used to compare the fault tolerance of different architectures and to design a regularizer improving the fault toleranc e of a given architecture. We design an algorithm achieving fault tolerance usin g a reasonable number of neurons. In addition to the theoretical proof, we also provide experimental validation of our results and suggest a connection to the g eneralization capacity problem.
****

Variational Hashing-based Collaborative Filtering with Self-Masking
Casper Hansen,Christian Hansen,Jakob Grue Simonsen,Stephen Alstrup,Christina Lio ma
Hashing-based collaborative filtering learns binary vector representations (hash codes) of users and items, such that recommendations can be computed very effic iently using the Hamming distance, which is simply the sum of differing bits bet ween two hash codes. A problem with hashing-based collaborative filtering using the Hamming distance, is that each bit is equally weighted in the distance compu tation, but in practice some bits might encode more important properties than ot her bits, where the importance depends on the user.
To this end, we propose an end-to-end trainable variational hashing-based collab orative filtering approach that uses the novel concept of self-masking: the user hash code acts as a mask on the items (using the Boolean AND operation), such t hat it learns to encode which bits are important to the user, rather than the us er's preference towards the underlying item property that the bits represent. Th is allows a binary user-level importance weighting of each item without the need to store additional weights for each user. We experimentally evaluate our appro ach against state-of-the-art baselines on 4 datasets, and obtain significant gai ns of up to 12% in NDCG. We also make available an efficient implementation of s elf-masking, which experimentally yields <4% runtime overhead compared to the st andard Hamming distance.
****

Neural Network Branching for Neural Network Verification
Jingyue Lu,M. Pawan Kumar
Formal verification of neural networks is essential for their deployment in safe ty-critical areas. Many available formal verification methods have been shown to be instances of a unified Branch and Bound (BaB) formulation. We propose a nove l framework for designing an effective branching strategy for BaB. Specifically, we learn a graph neural network (GNN) to imitate the strong branching heuristic behaviour. Our framework differs from previous methods for learning to branch i n two main aspects. Firstly, our framework directly treats the neural network we want to verify as a graph input for the GNN. Secondly, we develop an intuitive forward and backward embedding update schedule. Empirically, our framework achie ves roughly 50% reduction in both the number of branches and the time require d for verification on various convolutional networks when compared to the best a vailable hand-designed branching strategy. In addition, we show that our GNN mod el enjoys both horizontal and vertical transferability. Horizontally, the model trained on easy properties performs well on properties of increased difficulty l evels. Vertically, the model trained on small neural networks achieves similar p erformance on large neural networks.
****

SoftLoc: Robust Temporal Localization under Label Misalignment
Julien Schroeter,Kirill Sidorov,Dave Marshall
This work addresses the long-standing problem of robust event localization in th e presence of temporally of misaligned labels in the training data. We propose a novel versatile loss function that generalizes a number of training regimes fro m standard fully-supervised cross-entropy to count-based weakly-supervised learn ing. Unlike classical models which are constrained to strictly fit the annotatio ns during training, our soft localization learning approach relaxes the reliance

on the exact position of labels instead. Training with this new loss function exhibits strong robustness to temporal misalignment of labels, thus alleviating the burden of precise annotation of temporal sequences. We demonstrate state-of-the-art performance against standard benchmarks in a number of challenging experiments and further show that robustness to label noise is not achieved at the expense of raw performance.
****

VideoFlow: A Conditional Flow-Based Model for Stochastic Video Generation
Manoj Kumar,Mohammad Babaeizadeh,Dumitru Erhan,Chelsea Finn,Sergey Levine,Laurent Dinh,Durk Kingma
Generative models that can model and predict sequences of future events can, in principle, learn to capture complex real-world phenomena, such as physical interactions. However, a central challenge in video prediction is that the future is highly uncertain: a sequence of past observations of events can imply many possible futures. Although a number of recent works have studied probabilistic models that can represent uncertain futures, such models are either extremely expensive computationally as in the case of pixel-level autoregressive models, or do not directly optimize the likelihood of the data. To our knowledge, our work is the first to propose multi-frame video prediction with normalizing flows, which allows for direct optimization of the data likelihood, and produces high-quality stochastic predictions. We describe an approach for modeling the latent space dynamics, and demonstrate that flow-based generative models offer a viable and competitive approach to generative modeling of video.
****

Adaptive Data Augmentation with Deep Parallel Generative Models
Boli Fang,Miao Jiang,Abhirag Nagpure,Jerry Shen
Data augmentation(DA) is a useful technique to enlarge the size of the training set and prevent overfitting for different machine learning tasks when training data is scarce. However, current data augmentation techniques rely heavily on human design and domain knowledge, and existing automated approaches are yet to fully exploit the latent features in the training dataset. In this paper we propose an adaptive DA strategy based on generative models, where the training set adaptively enriches itself with sample images automatically constructed from deep generative models trained in parallel. We demonstrate by experiments that our data augmentation strategy, with little model-specific considerations, can be easily adapted to cross-domain deep learning/machine learning tasks such as image classification and image inpainting, while significantly improving model performance in both tasks.
****

Domain-invariant Learning using Adaptive Filter Decomposition
Ze Wang,Xiuyuan Cheng,Guillermo Sapiro,Qiang Qiu
Domain shifts are frequently encountered in real-world scenarios. In this paper, we consider the problem of domain-invariant deep learning by explicitly modeling domain shifts with only a small amount of domain-specific parameters in a Convolutional Neural Network (CNN). By exploiting the observation that a convolutional filter can be well approximated as a linear combination of a small set of basis elements, we show for the first time, both empirically and theoretically, that domain shifts can be effectively handled by decomposing a regular convolutional layer into a domain-specific basis layer and a domain-shared basis coefficient layer, while both remain convolutional. An input channel will now first convolve spatially only with each respective domain-specific basis to ``absorb" domain variations, and then output channels are linearly combined using common basis coefficients trained to promote shared semantics across domains. We use toy examples, rigorous analysis, and real-world examples to show the framework's effectiveness in cross-domain performance and domain adaptation. With the proposed architecture, we need only a small set of basis elements to model each additional domain, which brings a negligible amount of additional parameters, typically a few hundred.
****

Adversarial Policies: Attacking Deep Reinforcement Learning

Adam Gleave,Michael Dennis,Cody Wild,Neel Kant,Sergey Levine,Stuart Russell
Deep reinforcement learning (RL) policies are known to be vulnerable to adversarial perturbations to their observations, similar to adversarial examples for classifiers. However, an attacker is not usually able to directly modify another agent's observations. This might lead one to wonder: is it possible to attack an RL agent simply by choosing an adversarial policy acting in a multi-agent environment so as to create natural observations that are adversarial? We demonstrate the existence of adversarial policies in zero-sum games between simulated humanoid robots with proprioceptive observations, against state-of-the-art victims trained via self-play to be robust to opponents. The adversarial policies reliably win against the victims but generate seemingly random and uncoordinated behavior. We find that these policies are more successful in high-dimensional environments, and induce substantially different activations in the victim policy network than when the victim plays against a normal opponent. Videos are available at https://adversarialpolicies.github.io/.
****
Escaping Saddle Points Faster with Stochastic Momentum
Jun-Kun Wang,Chi-Heng Lin,Jacob Abernethy
Stochastic gradient descent (SGD) with stochastic momentum is popular in nonconvex stochastic optimization and particularly for the training of deep neural networks. In standard SGD, parameters are updated by improving along the path of the gradient at the current iterate on a batch of examples, where the addition of a ``momentum'' term biases the update in the direction of the previous change in parameters. In non-stochastic convex optimization one can show that a momentum adjustment provably reduces convergence time in many settings, yet such results have been elusive in the stochastic and non-convex settings. At the same time, a widely-observed empirical phenomenon is that in training deep networks stochastic momentum appears to significantly improve convergence time, variants of it have flourished in the development of other popular update methods, e.g. ADAM, AMSGrad, etc. Yet theoretical justification for the use of stochastic momentum has remained a significant open question. In this paper we propose an answer: stochastic momentum improves deep network training because it modifies SGD to escape saddle points faster and, consequently, to more quickly find a second order stationary point. Our theoretical results also shed light on the related question of how to choose the ideal momentum parameter--our analysis suggests that $\beta \in [0,1)$ should be large (close to 1), which comports with empirical findings. We also provide experimental findings that further validate these conclusions.
****
Few-shot Text Classification with Distributional Signatures
Yujia Bao,Menghua Wu,Shiyu Chang,Regina Barzilay
In this paper, we explore meta-learning for few-shot text classification. Meta-learning has shown strong performance in computer vision, where low-level patterns are transferable across learning tasks. However, directly applying this approach to text is challenging--lexical features highly informative for one task may be insignificant for another. Thus, rather than learning solely from words, our model also leverages their distributional signatures, which encode pertinent word occurrence patterns. Our model is trained within a meta-learning framework to map these signatures into attention scores, which are then used to weight the lexical representations of words. We demonstrate that our model consistently outperforms prototypical networks learned on lexical knowledge (Snell et al., 2017) in both few-shot text classification and relation classification by a significant margin across six benchmark datasets (20.0% on average in 1-shot classification).
****
RotationOut as a Regularization Method for Neural Network
Kai Hu,Barnabas Poczos
In this paper, we propose a novel regularization method, RotationOut, for neural networks.
Different from Dropout that handles each neuron/channel independently, RotationOut regards its input layer as an entire vector and introduces regularization by

randomly rotating the vector.
RotationOut can also be used in convolutional layers and recurrent layers with a small modification.
We further use a noise analysis method to interpret the difference between RotationOut and Dropout in co-adaptation reduction.
Using this method, we also show how to use RotationOut/Dropout together with Batch Normalization.
Extensive experiments in vision and language tasks are conducted to show the effectiveness of the proposed method.
Codes will be available.
****

Universal Approximation with Deep Narrow Networks
Patrick Kidger,Terry Lyons
The classical Universal Approximation Theorem certifies that the universal approximation property holds for the class of neural networks of arbitrary width. Here we consider the natural `dual' theorem for width-bounded networks of arbitrary depth. Precisely, let n be the number of inputs neurons, m be the number of output neurons, and let ρ be any nonaffine continuous function, with a continuous nonzero derivative at some point. Then we show that the class of neural networks of arbitrary depth, width $n + m + 2$, and activation function ρ, exhibits the universal approximation property with respect to the uniform norm on compact subsets of $\mathbb{R}^n$. This covers every activation function possible to use in practice; in particular this includes polynomial activation functions, making this genuinely different to the classical case. We go on to consider extensions of this result. First we show an analogous result for a certain class of nowhere differentiable activation functions. Second we establish an analogous result for noncompact domains, by showing that deep narrow networks with the ReLU activation function exhibit the universal approximation property with respect to the p-norm on $\mathbb{R}^n$. Finally we show that width of only $n + m + 1$ suffices for `most' activation functions.
****

A Dynamic Approach to Accelerate Deep Learning Training
John Osorio,Adrià Armejach,Eric Petit,Marc Casas
Mixed-precision arithmetic combining both single- and half-precision operands in the same operation have been successfully applied to train deep neural networks. Despite the advantages of mixed-precision arithmetic in terms of reducing the need for key resources like memory bandwidth or register file size, it has a limited capacity for diminishing computing costs and requires 32 bits to represent its output operands. This paper proposes two approaches to replace mixed-precision for half-precision arithmetic during a large portion of the training. The first approach achieves accuracy ratios slightly slower than the state-of-the-art by using half-precision arithmetic during more than 99% of training. The second approach reaches the same accuracy as the state-of-the-art by dynamically switching between half- and mixed-precision arithmetic during training. It uses half-precision during more than 94% of the training process. This paper is the first in demonstrating that half-precision can be used for a very large portion of DNNs training and still reach state-of-the-art accuracy.
****

Geometric Insights into the Convergence of Nonlinear TD Learning
David Brandfonbrener,Joan Bruna
While there are convergence guarantees for temporal difference (TD) learning when using linear function approximators, the situation for nonlinear models is far less understood, and divergent examples are known. Here we take a first step towards extending theoretical convergence guarantees to TD learning with nonlinear function approximation. More precisely, we consider the expected learning dynamics of the TD(0) algorithm for value estimation. As the step-size converges to zero, these dynamics are defined by a nonlinear ODE which depends on the geometry of the space of function approximators, the structure of the underlying Markov chain, and their interaction. We find a set of function approximators that includes ReLU networks and has geometry amenable to TD learning regardless of environ

ment, so that the solution performs about as well as linear TD in the worst case . Then, we show how environments that are more reversible induce dynamics that a re better for TD learning and prove global convergence to the true value functio n for well-conditioned function approximators. Finally, we generalize a divergen t counterexample to a family of divergent problems to demonstrate how the intera ction between approximator and environment can go wrong and to motivate the assu mptions needed to prove convergence.
****
Efficient Multivariate Bandit Algorithm with Path Planning
Keyu Nie,Zezhong Zhang,Ted Tao Yuan,Rong Song,Pauline Berry Burke
In this paper, we solve the arms exponential exploding issues in multivariate Mu lti-Armed Bandit (Multivariate-MAB) problem when the arm dimension hierarchy is considered. We propose a framework called path planning (TS-PP) which utilizes d ecision graph/trees to model arm reward success rate with m-way dimension intera ction, and adopts Thompson sampling (TS) for heuristic search of arm selection. Naturally, it is quite straightforward to combat the curse of dimensionality usi ng a serial processes that operates sequentially by focusing on one dimension pe r each process. For our best acknowledge, we are the first to solve Multivariat e-MAB problem using graph path planning strategy and deploying alike Monte-Carlo tree search ideas. Our proposed method utilizing tree models has advantages com paring with traditional models such as general linear regression. Simulation stu dies validate our claim by achieving faster convergence speed, better efficient optimal arm allocation and lower cumulative regret.
****
Learning Self-Correctable Policies and Value Functions from Demonstrations with Negative Sampling
Yuping Luo,Huazhe Xu,Tengyu Ma
Imitation learning, followed by reinforcement learning algorithms, is a promisin g paradigm to solve complex control tasks sample-efficiently. However, learning from demonstrations often suffers from the covariate shift problem, which result s
in cascading errors of the learned policy. We introduce a notion of conservative ly extrapolated value functions, which provably lead to policies with self-corre ction. We design an algorithm Value Iteration with Negative Sampling (VINS) that practically learns such value functions with conservative extrapolation. We sho w that VINS can correct mistakes of the behavioral cloning policy on simulated r obotics benchmark tasks. We also propose the algorithm of using VINS to initiali ze a reinforcement learning algorithm, which is shown to outperform prior works in sample efficiency.
****
Exploring Model-based Planning with Policy Networks
Tingwu Wang,Jimmy Ba
Model-based reinforcement learning (MBRL) with model-predictive control or online planning has shown great potential for locomotion control tasks in both sample efficiency and asymptotic performance. Despite the successes, the existin g
planning methods search from candidate sequences randomly generated in the action space, which is inefficient in complex high-dimensional environments. In this paper, we propose a novel MBRL algorithm, model-based policy planning (POPLIN), that combines policy networks with online planning. More specifically, we formulate action planning at each time-step as an optimization problem using neural networks. We experiment with both optimization w.r.t. the action sequence s
initialized from the policy network, and also online optimization directly w.r.t . the
parameters of the policy network. We show that POPLIN obtains state-of-the-art performance in the MuJoCo benchmarking environments, being about 3x more sample efficient than the state-of-the-art algorithms, such as PETS, TD3 and SAC .
To explain the effectiveness of our algorithm, we show that the optimization sur

face
in parameter space is smoother than in action space. Further more, we found the distilled policy network can be effectively applied without the expansive model predictive control during test time for some environments such as Cheetah. Code is released.
****
Benchmarking Model-Based Reinforcement Learning
Tingwu Wang,Xuchan Bao,Ignasi Clavera,Jerrick Hoang,Yeming Wen,Eric Langlois,Shunshi Zhang,Guodong Zhang,Pieter Abbeel,Jimmy Ba
Model-based reinforcement learning (MBRL) is widely seen as having the potential to be significantly more sample efficient than model-free RL. However, research in model-based RL has not been very standardized. It is fairly common for authors to experiment with self-designed environments, and there are several separate lines of research, which are sometimes closed-sourced or not reproducible. Accordingly, it is an open question how these various existing algorithms perform relative to each other. To facilitate research in MBRL, in this paper we gather a wide collection of MBRL algorithms and propose over 18 benchmarking environments specially designed for MBRL. We benchmark these algorithms with unified problem settings, including noisy environments. Beyond cataloguing performance, we explore and unify the underlying algorithmic differences across MBRL algorithms. We characterize three key research challenges for future MBRL research: the dynamics bottleneck, the planning horizon dilemma, and the early-termination dilemma. Finally, to facilitate future research on MBRL, we open-source our benchmark.
****
Encoder-decoder Network as Loss Function for Summarization
Glen Jeh
We present a new approach to defining a sequence loss function to train a summarizer by using a secondary encoder-decoder as a loss function, alleviating a shortcoming of word level training for sequence outputs. The technique is based on the intuition that if a summary is a good one, it should contain the most essential information from the original article, and therefore should itself be a good input sequence, in lieu of the original, from which a summary can be generated. We present experimental results where we apply this additional loss function to a general abstractive summarizer on a news summarization dataset. The result is an improvement in the ROUGE metric and an especially large improvement in human evaluations, suggesting enhanced performance that is competitive with specialized state-of-the-art models.
****
On Identifiability in Transformers
Gino Brunner,Yang Liu,Damian Pascual,Oliver Richter,Massimiliano Ciaramita,Roger Wattenhofer
In this paper we delve deep in the Transformer architecture by investigating two of its core components: self-attention and contextual embeddings. In particular, we study the identifiability of attention weights and token embeddings, and the aggregation of context into hidden tokens. We show that, for sequences longer than the attention head dimension, attention weights are not identifiable. We propose effective attention as a complementary tool for improving explanatory interpretations based on attention. Furthermore, we show that input tokens retain to a large degree their identity across the model. We also find evidence suggesting that identity information is mainly encoded in the angle of the embeddings and gradually decreases with depth. Finally, we demonstrate strong mixing of input information in the generation of contextual embeddings by means of a novel quantification method based on gradient attribution. Overall, we show that self-attention distributions are not directly interpretable and present tools to better un

derstand and further investigate Transformer models.
****
Automated curriculum generation through setter-solver interactions
Sebastien Racaniere,Andrew Lampinen,Adam Santoro,David Reichert,Vlad Firoiu,Timothy Lillicrap
Reinforcement learning algorithms use correlations between policies and rewards to improve agent performance. But in dynamic or sparsely rewarding environments these correlations are often too small, or rewarding events are too infrequent to make learning feasible. Human education instead relies on curricula –the breakdown of tasks into simpler, static challenges with dense rewards– to build up to complex behaviors. While curricula are also useful for artificial agents, hand-crafting them is time consuming. This has lead researchers to explore automatic curriculum generation. Here we explore automatic curriculum generation in rich,dynamic environments. Using a setter-solver paradigm we show the importance of considering goal validity, goal feasibility, and goal coverage to construct useful curricula. We demonstrate the success of our approach in rich but sparsely rewarding 2D and 3D environments, where an agent is tasked to achieve a single goal selected from a set of possible goals that varies between episodes, and identify challenges for future work. Finally, we demonstrate the value of a novel technique that guides agents towards a desired goal distribution. Altogether, these results represent a substantial step towards applying automatic task curricula to learn complex, otherwise unlearnable goals, and to our knowledge are the first to demonstrate automated curriculum generation for goal-conditioned agents in environments where the possible goals vary between episodes.
****
Deep Multi-View Learning via Task-Optimal CCA
Heather D. Couture,Roland Kwitt,J.S. Marron,Melissa Troester,Charles M. Perou,Marc Niethammer
Canonical Correlation Analysis (CCA) is widely used for multimodal data analysis and, more recently, for discriminative tasks such as multi-view learning; however, it makes no use of class labels. Recent CCA methods have started to address this weakness but are limited in that they do not simultaneously optimize the CCA projection for discrimination and the CCA projection itself, or they are linear only. We address these deficiencies by simultaneously optimizing a CCA-based and a task objective in an end-to-end manner. Together, these two objectives learn a non-linear CCA projection to a shared latent space that is highly correlated and discriminative. Our method shows a significant improvement over previous state-of-the-art (including deep supervised approaches) for cross-view classification (8.5% increase), regularization with a second view during training when only one view is available at test time (2.2-3.2%), and semi-supervised learning (15%) on real data.
****
Bandlimiting Neural Networks Against Adversarial Attacks
Yuping Lin,Kasra Ahmadi K. A.,Hui Jiang
In this paper, we study the adversarial attack and defence problem in deep learning from the perspective of Fourier analysis. We first explicitly compute the Fourier transform of deep ReLU neural networks and show that there exist decaying but non-zero high frequency components in the Fourier spectrum of neural networks. We then demonstrate that the vulnerability of neural networks towards adversarial samples can be attributed to these insignificant but non-zero high frequency components. Based on this analysis, we propose to use a simple post-averaging technique to smooth out these high frequency components to improve the robustness of neural networks against adversarial attacks. Experimental results on the ImageNet and the CIFAR-10 datasets have shown that our proposed method is universally effective to defend many existing adversarial attacking methods proposed in the literature, including FGSM, PGD, DeepFool and C&W attacks. Our post-averaging method is simple since it does not require any re-training, and meanwhile it can successfully defend over 80-96% of the adversarial samples generated by these methods without introducing significant performance degradation (less than 2%) on the original clean images.

****
Progressive Memory Banks for Incremental Domain Adaptation
Nabiha Asghar,Lili Mou,Kira A. Selby,Kevin D. Pantasdo,Pascal Poupart,Xin Jiang
This paper addresses the problem of incremental domain adaptation (IDA) in natural language processing (NLP). We assume each domain comes one after another, and that we could only access data in the current domain. The goal of IDA is to build a unified model performing well on all the domains that we have encountered. We adopt the recurrent neural network (RNN) widely used in NLP, but augment it with a directly parameterized memory bank, which is retrieved by an attention mechanism at each step of RNN transition. The memory bank provides a natural way of IDA: when adapting our model to a new domain, we progressively add new slots to the memory bank, which increases the number of parameters, and thus the model capacity. We learn the new memory slots and fine-tune existing parameters by back-propagation. Experimental results show that our approach achieves significantly better performance than fine-tuning alone. Compared with expanding hidden states, our approach is more robust for old domains, shown by both empirical and theoretical results. Our model also outperforms previous work of IDA including elastic weight consolidation and progressive neural networks in the experiments.
****
MMD GAN with Random-Forest Kernels
Tao Huang,Zhen Han,Xu Jia,Hanyuan Hang
In this paper, we propose a novel kind of kernel, random forest kernel, to enhance the empirical performance of MMD GAN. Different from common forests with deterministic routings, a probabilistic routing variant is used in our innovated random-forest kernel, which is possible to merge with the CNN frameworks. Our proposed random-forest kernel has the following advantages: From the perspective of random forest, the output of GAN discriminator can be viewed as feature inputs to the forest, where each tree gets access to merely a fraction of the features, and thus the entire forest benefits from ensemble learning. In the aspect of kernel method, random-forest kernel is proved to be characteristic, and therefore suitable for the MMD structure. Besides, being an asymmetric kernel, our random-forest kernel is much more flexible, in terms of capturing the differences between distributions. Sharing the advantages of CNN, kernel method, and ensemble learning, our random-forest kernel based MMD GAN obtains desirable empirical performances on CIFAR-10, CelebA and LSUN bedroom data sets. Furthermore, for the sake of completeness, we also put forward comprehensive theoretical analysis to support our experimental results.
****
What graph neural networks cannot learn: depth vs width
Andreas Loukas
This paper studies the expressive power of graph neural networks falling within the message-passing framework (GNNmp). Two results are presented. First, GNNmp are shown to be Turing universal under sufficient conditions on their depth, width, node attributes, and layer expressiveness. Second, it is discovered that GNNmp can lose a significant portion of their power when their depth and width is restricted. The proposed impossibility statements stem from a new technique that enables the repurposing of seminal results from distributed computing and leads to lower bounds for an array of decision, optimization, and estimation problems involving graphs. Strikingly, several of these problems are deemed impossible unless the product of a GNNmp's depth and width exceeds a polynomial of the graph size; this dependence remains significant even for tasks that appear simple or when considering approximation.
****
INFERENCE, PREDICTION, AND ENTROPY RATE OF CONTINUOUS-TIME, DISCRETE-EVENT PROCESSES
Sarah Marzen,James P. Crutchfield
The inference of models, prediction of future symbols, and entropy rate estimation of discrete-time, discrete-event processes is well-worn ground. However, many time series are better conceptualized as continuous-time, discrete-event processes. Here, we provide new methods for inferring models, predicting future symbol

s, and estimating the entropy rate of continuous-time, discrete-event processes. The methods rely on an extension of Bayesian structural inference that takes advantage of neural network’s universal approximation power. Based on experiments with simple synthetic data, these new methods seem to be competitive with state-of- the-art methods for prediction and entropy rate estimation as long as the correct model is inferred.
****

RTFM: Generalising to New Environment Dynamics via Reading
Victor Zhong,Tim Rocktäschel,Edward Grefenstette
Obtaining policies that can generalise to new environments in reinforcement learning is challenging. In this work, we demonstrate that language understanding via a reading policy learner is a promising vehicle for generalisation to new environments. We propose a grounded policy learning problem, Read to Fight Monsters (RTFM), in which the agent must jointly reason over a language goal, relevant dynamics described in a document, and environment observations. We procedurally generate environment dynamics and corresponding language descriptions of the dynamics, such that agents must read to understand new environment dynamics instead of memorising any particular information. In addition, we propose txt2π, a model that captures three-way interactions between the goal, document, and observations. On RTFM, txt2π generalises to new environments with dynamics not seen during training via reading. Furthermore, our model outperforms baselines such as FiLM and language-conditioned CNNs on RTFM. Through curriculum learning, txt2π produces policies that excel on complex RTFM tasks requiring several reasoning and coreference steps.
****

MIM: Mutual Information Machine
Micha Livne,Kevin Swersky,David J. Fleet
We introduce the Mutual Information Machine (MIM), an autoencoder framework
for learning joint distributions over observations and latent states.
The model formulation reflects two key design principles: 1) symmetry, to encourage
the encoder and decoder to learn different factorizations of the same
underlying distribution; and 2) mutual information, to encourage the learning
of useful representations for downstream tasks.
The objective comprises the Jensen-Shannon divergence between the encoding and
decoding joint distributions, plus a mutual information regularizer.
We show that this can be bounded by a tractable cross-entropy loss between
the true model and a parameterized approximation, and relate this to
maximum likelihood estimation and variational autoencoders.
Experiments show that MIM is capable of learning a latent representation with high mutual information,
and good unsupervised clustering, while providing NLL comparable to VAE
(with a sufficiently expressive architecture).
****

Constant Time Graph Neural Networks
Ryoma Sato,Makoto Yamada,Hisashi Kashima
The recent advancements in graph neural networks (GNNs) have led to state-of-the-art performances in various applications, including chemo-informatics, question-answering systems, and recommender systems. However, scaling up these methods to huge graphs such as social network graphs and web graphs still remains a challenge. In particular, the existing methods for accelerating GNNs are either not theoretically guaranteed in terms of approximation error, or they require at least a linear time computation cost.
In this study, we analyze the neighbor sampling technique to obtain a constant time approximation algorithm for GraphSAGE, the graph attention networks (GAT), and the graph convolutional networks (GCN). The proposed approximation algorithm can theoretically guarantee the precision of approximation. The key advantage of the proposed approximation algorithm is that the complexity is completely indep

endent of the numbers of the nodes, edges, and neighbors of the input and depend s only on the error tolerance and confidence probability. To the best of our kno wledge, this is the first constant time approximation algorithm for GNNs with a theoretical guarantee. Through experiments using synthetic and real-world datase ts, we demonstrate the speed and precision of the proposed approximation algorit hm and validate our theoretical results.
****
AutoLR: A Method for Automatic Tuning of Learning Rate
Nipun Kwatra,V Thejas,Nikhil Iyer,Ramachandran Ramjee,Muthian Sivathanu
One very important hyperparameter for training deep neural networks is the learning rate of the optimizer. The choice of learning rate schedule determines the computational cost of getting close to a minima, how close you actually get to the minima, and most importantly the kind of local minima (wide/narrow) attained. The kind of minima attained has a significant impact on the generalization accuracy of the network. Current systems employ hand tuned learning rate schedules, which are painstakingly tuned for each network and dataset. Given that the state space of schedules is huge, finding a satisfactory learning rate schedule can be very time consuming. In this paper, we present AutoLR, a method for auto-tuning the learning rate as training proceeds. Our method works with any optimizer, and we demonstrate results on SGD, Momentum, and Adam optimizers.

We extensively evaluate AutoLR on multiple datasets, models, and across multiple optimizers. We compare favorably against state of the art learning rate schedules for the given dataset and models, including for ImageNet on Resnet-50, Cifar-10 on Resnet-18, and SQuAD fine-tuning on BERT. For example, AutoLR achieves an EM score of 81.2 on SQuAD v1.1 with BERT_BASE compared to 80.8 reported in (Devlin et al. (2018)) by just auto-tuning the learning rate schedule. To the best of our knowledge, this is the first automatic learning rate tuning scheme to achieve state of the art generalization accuracy on these datasets with the given models.

****
Generating Robust Audio Adversarial Examples using Iterative Proportional Clippi ng
Hongting Zhang,Qiben Yan,Pan Zhou
Audio adversarial examples, imperceptible to humans, have been constructed to at tack automatic speech recognition (ASR) systems. However, the adversarial exampl es generated by existing approaches usually involve notable noise, especially du ring the periods of silence and pauses, which may lead to the detection of such attacks. This paper proposes a new approach to generate adversarial audios using Iterative Proportional Clipping (IPC), which exploits temporal dependency in or iginal audios to significantly limit human-perceptible noise. Specifically, in e very iteration of optimization, we use a backpropagation model to learn the raw perturbation on the original audio to construct our clipping. We then impose a c onstraint on the perturbation at the positions with lower sound intensity across the time domain to eliminate the perceptible noise during the silent periods or pauses. IPC preserves the linear proportionality between the original audio and the perturbed one to maintain the temporal dependency. We show that the propose d approach can successfully attack the latest state-of-the-art ASR model Wav2let ter+, and only requires a few minutes to generate an audio adversarial example. Experimental results also demonstrate that our approach succeeds in preserving t emporal dependency and can bypass temporal dependency based defense mechanisms.
****
Optimal Attacks on Reinforcement Learning Policies
Alessio Russo,Alexandre Proutiere
Control policies, trained using the Deep Reinforcement Learning, have been recen tly shown to be vulnerable to adversarial attacks introducing even very small pe rturbations to the policy input. The attacks proposed so far have been designed using heuristics, and build on existing adversarial example crafting techniques

used to dupe classifiers in supervised learning. In contrast, this paper investi
gates the problem of devising optimal attacks, depending on a well-defined attac
ker's objective, e.g., to minimize the main agent average reward. When the polic
y and the system dynamics, as well as rewards, are known to the attacker, a scen
ario referred to as a white-box attack, designing optimal attacks amounts to sol
ving a Markov Decision Process. For what we call black-box attacks, where neithe
r the policy nor the system is known, optimal attacks can be trained using Reinf
orcement Learning techniques. Through numerical experiments, we demonstrate the
efficiency of our attacks compared to existing attacks (usually based on Gradie
nt methods). We further quantify the potential impact of attacks and establish i
ts connection to the smoothness of the policy under attack. Smooth policies are
naturally less prone to attacks (this explains why Lipschitz policies, with resp
ect to the state, are more resilient). Finally, we show that from the main agent
perspective, the system uncertainties and the attacker can be modelled as a Par
tially Observable Markov Decision Process. We actually demonstrate that using Re
inforcement Learning techniques tailored to POMDP (e.g. using Recurrent Neural N
etworks) leads to more resilient policies.
****

Multi-Agent Hierarchical Reinforcement Learning for Humanoid Navigation
Glen Berseth,Brandon haworth,Seonghyeon Moon,Mubbasir Kapadia,Petros Faloutsos
Multi-agent reinforcement learning is a particularly challenging problem. Curren
t
methods have made progress on cooperative and competitive environments with
particle-based agents. Little progress has been made on solutions that could op-
erate in the real world with interaction, dynamics, and humanoid robots. In this
work, we make a significant step in multi-agent models on simulated humanoid
robot navigation by combining Multi-Agent Reinforcement Learning (MARL)
with Hierarchical Reinforcement Learning (HRL). We build on top of fou
nda-
tional prior work in learning low-level physical controllers for locomotion
and
add a layer to learn decentralized policies for multi-agent goal-directed collis
ion
avoidance systems. A video of our results on a multi-agent pursuit environment
can be seen here

****

SMiRL: Surprise Minimizing RL in Entropic Environments
Glen Berseth,Daniel Geng,Coline Devin,Dinesh Jayaraman,Chelsea Finn,Sergey Levin
e
All living organisms struggle against the forces of nature to carve out niches w
here
they can maintain relative stasis. We propose that such a search for order amids
t
chaos might offer a unifying principle for the emergence of useful behaviors in
artificial agents. We formalize this idea into an unsupervised reinforcement lea
rning
method called surprise minimizing RL (SMiRL). SMiRL trains an agent with the
objective of maximizing the probability of observed states under a model trained
on
all previously seen states. The resulting agents acquire several proactive behav
iors
to seek and maintain stable states such as balancing and damage avoidance, that
are closely tied to the affordances of the environment and its prevailing source
s
of entropy, such as winds, earthquakes, and other agents. We demonstrate that
our surprise minimizing agents can successfully play Tetris, Doom, and control
a humanoid to avoid falls, without any task-specific reward supervision
. We
further show that SMiRL can be used as an unsupervised pre-training objective

that substantially accelerates subsequent reward-driven learning
****
Mesh-Free Unsupervised Learning-Based PDE Solver of Forward and Inverse problems
Leah Bar,Nir Sochen
We introduce a novel neural network-based partial differential equations solver for forward and inverse problems. The solver is grid free, mesh free and shape free, and the solution is approximated by a neural network.
We employ an unsupervised approach such that the input to the network is a points set in an arbitrary domain, and the output is the
set of the corresponding function values. The network is trained to minimize deviations of the learned function from the PDE solution and
satisfy the boundary conditions.
The resulting solution in turn is an explicit smooth differentiable function with a known analytical form.

Unlike other numerical methods such as finite differences and finite elements, the derivatives of the desired function can be analytically calculated to any order. This framework therefore, enables the solution of high order non-linear PDEs. The proposed algorithm is a unified formulation of both forward and inverse problems
where the optimized loss function consists of few elements: fidelity terms of L2 and L infinity norms, boundary conditions constraints and additional regularizers. This setting is flexible in the sense that regularizers can be tailored to specific
problems. We demonstrate our method on a free shape 2D second order elliptical system with application to Electrical Impedance Tomography (EIT).
****
Input Complexity and Out-of-distribution Detection with Likelihood-based Generative Models
Joan Serrà,David Álvarez,Vicenç Gómez,Olga Slizovskaia,José F. Núñez,Jordi Luque
Likelihood-based generative models are a promising resource to detect out-of-distribution (OOD) inputs which could compromise the robustness or reliability of a machine learning system. However, likelihoods derived from such models have been shown to be problematic for detecting certain types of inputs that significantly differ from training data. In this paper, we pose that this problem is due to the excessive influence that input complexity has in generative models' likelihoods. We report a set of experiments supporting this hypothesis, and use an estimate of input complexity to derive an efficient and parameter-free OOD score, which can be seen as a likelihood-ratio, akin to Bayesian model comparison. We find such score to perform comparably to, or even better than, existing OOD detection approaches under a wide range of data sets, models, model sizes, and complexity estimates.
****
Sparse and Structured Visual Attention
Pedro Henrique Martins,Vlad Niculae,Zita Marinho,André F.T. Martins
Visual attention mechanisms have been widely used in image captioning models. In this paper, to better link the image structure with the generated text, we replace the traditional softmax attention mechanism by two alternative sparsity-promoting transformations: sparsemax and Total-Variation Sparse Attention (TVmax). With sparsemax, we obtain sparse attention weights, selecting relevant features. In order to promote sparsity and encourage fusing of the related adjacent spatial locations, we propose TVmax. By selecting relevant groups of features, the TVmax transformation improves interpretability. We present results in the Microsoft COCO and Flickr30k datasets, obtaining gains in comparison to softmax. TVmax outperforms the other compared attention mechanisms in terms of human-rated caption quality and attention relevance.
****
Network Pruning for Low-Rank Binary Index
Dongsoo Lee,Se Jung Kwon,Byeongwook Kim,Parichay Kapoor,Gu-Yeon Wei
  Pruning is an efficient model compression technique to remove redundancy in th

e connectivity of deep neural networks (DNNs). A critical problem to represent s parse matrices after pruning is that if fewer bits are used for quantization and pruning rate is enhanced, then the amount of index becomes relatively larger. M oreover, an irregular index form leads to low parallelism for convolutions and m atrix multiplications. In this paper, we propose a new network pruning technique that generates a low-rank binary index matrix to compress index data significan tly. Specifically, the proposed compression method finds a particular fine-grain ed pruning mask that can be decomposed into two binary matrices while decompress ing index data is performed by simple binary matrix multiplication. We also prop ose a tile-based factorization technique that not only lowers memory requirement s but also enhances compression ratio. Various DNN models (including conv layers and LSTM layers) can be pruned with much fewer indices compared to previous spa rse matrix formats while maintaining the same pruning rate.

****

Style-based Encoder Pre-training for Multi-modal Image Synthesis

Moustafa Meshry,Yixuan Ren,Ricardo Martin-Brualla,Larry Davis,Abhinav Shrivastav a

Image-to-image (I2I) translation aims to translate images from one domain to ano ther. To tackle the multi-modal version of I2I translation, where input and outp ut domains have a one-to-many relation, an extra latent input is provided to the generator to specify a particular output. Recent works propose involved trainin g objectives to learn a latent embedding, jointly with the generator, that model s the distribution of possible outputs. Alternatively, we study a simple, yet po werful pre-training strategy for multi-modal I2I translation. We first pre-train an encoder, using a proxy task, to encode the style of an image, such as color and texture, into a low-dimensional latent style vector. Then we train a generat or to transform an input image along with a style-code to the output domain. Our generator achieves state-of-the-art results on several benchmarks with a traini ng objective that includes just a GAN loss and a reconstruction loss, which simp lifies and speeds up the training significantly compared to competing approaches . We further study the contribution of different loss terms to learning the task of multi-modal I2I translation, and finally we show that the learned style embe dding is not dependent on the target domain and generalizes well to other domain s.

****

LDMGAN: Reducing Mode Collapse in GANs with Latent Distribution Matching

Zhiwen Zuo,Lei Zhao,Huiming Zhang,Qihang Mo,Haibo Chen,Zhizhong Wang,AiLin Li,Li hong Qiu,Wei Xing,Dongming Lu

Generative Adversarial Networks (GANs) have shown impressive results in modeling distributions over complicated manifolds such as those of natural images. Howev er, GANs often suffer from mode collapse, which means they are prone to characte rize only a single or a few modes of the data distribution. In order to address this problem, we propose a novel framework called LDMGAN. We ■rst introduce Late nt Distribution Matching (LDM) constraint which regularizes the generator by ali gning distribution of generated samples with that of real samples in latent spac e. To make use of such latent space, we propose a regularized AutoEncoder (AE) t hat maps the data distribution to prior distribution in encoded space. Extensive experiments on synthetic data and real world datasets show that our proposed fr amework signi■cantly improves GAN’s stability and diversity.

****

Bootstrapping the Expressivity with Model-based Planning

Kefan Dong,Yuping Luo,Tengyu Ma

We compare the model-free reinforcement learning with the model-based approaches through the lens of the expressive power of neural networks for policies, Q-f unctions, and dynamics. We show, theoretically and empirically, that even for o ne-dimensional continuous state space, there are many MDPs whose optimal Q-fun ctions and policies are much more complex than the dynamics. We hypothesize many real-world MDPs also have a similar property. For these MDPs, model-based plann ing is a favorable algorithm, because the resulting policies can approximate the optimal policy significantly better than a neural network parameterization can,

and model-free or model-based policy optimization rely on policy parameterizati
on. Motivated by the theory, we apply a simple multi-step model-based bootstrapp
ing planner (BOOTS) to bootstrap a weak Q-function into a stronger policy. Emp
irical results show that applying BOOTS on top of model-based or model-free poli
cy optimization algorithms at the test time improves the performance on MuJoCo b
enchmark tasks.
****

DeepAGREL: Biologically plausible deep learning via direct reinforcement
Isabella Pozzi,Sander M. Bohte,Pieter R. Roelfsema
While much recent work has focused on biologically plausible variants of error-b
ackpropagation, learning in the brain seems to mostly adhere to a reinforcement
learning paradigm; biologically plausible neural reinforcement learning framewor
ks, however, were limited to shallow networks learning from compact and abstract
sensory representations. Here, we show that it is possible to generalize such a
pproaches to deep networks with an arbitrary number of layers.
We demonstrate the learning scheme - DeepAGREL - on classical and hard image-cla
ssification benchmarks requiring deep networks, namely MNIST, CIFAR10, and CIFAR
100, cast as direct reward tasks, both for deep fully connected, convolutional a
nd locally connected architectures. We show that for these tasks, DeepAGREL achi
eves an accuracy that is equal to supervised error-backpropagation, and the tria
l-and-error nature of such learning imposes only a very limited cost in terms of
training time. Thus, our results provide new insights into how deep learning ma
y be implemented in the brain.
****

Homogeneous Linear Inequality Constraints for Neural Network Activations
Thomas Frerix,Matthias Nießner,Daniel Cremers
We propose a method to impose homogeneous linear inequality constraints of the f
orm $Ax\leq 0$ on neural network activations. The proposed method allows a data-
driven training approach to be combined with modeling prior knowledge about the
task. One way to achieve this task is by means of a projection step at test time
after unconstrained training.
However, this is an expensive operation. By directly incorporating the constrain
ts into the architecture, we can significantly speed-up inference at test time;
for instance, our experiments show a speed-up of up to two orders of magnitude o
ver a projection method. Our algorithm computes a suitable parameterization of t
he feasible set at initialization and uses standard variants of stochastic gradi
ent descent to find solutions to the constrained network. Thus, the modeling con
straints are always satisfied during training. Crucially, our approach avoids to
solve an optimization problem at each training step or to manually trade-off da
ta and constraint fidelity with additional hyperparameters. We consider constrai
ned generative modeling as an important application domain and experimentally de
monstrate the proposed method by constraining a variational autoencoder.
****

Leveraging Simple Model Predictions for Enhancing its Performance
Amit Dhurandhar,Karthikeyan Shanmugam,Ronny Luss
There has been recent interest in improving performance of simple models for mul
tiple reasons such as interpretability, robust learning from small data, deploym
ent in memory constrained settings as well as environmental considerations. In t
his paper, we propose a novel method SRatio that can utilize information from hi
gh performing complex models (viz. deep neural networks, boosted trees, random f
orests) to reweight a training dataset for a potentially low performing simple m
odel such as a decision tree or a shallow network enhancing its performance. Our
method also leverages the per sample hardness estimate of the simple model whic
h is not the case with the prior works which primarily consider the complex mode
l's confidences/predictions and is thus conceptually novel. Moreover, we general
ize and formalize the concept of attaching probes to intermediate layers of a ne
ural network, which was one of the main ideas in previous work \citep{profweight
}, to other commonly used classifiers and incorporate this into our method. The
benefit of these contributions is witnessed in the experiments where on 6 UCI da
tasets and CIFAR-10 we outperform competitors in a majority (16 out of 27) of th

e cases and tie for best performance in the remaining cases. In fact, in a coupl e of cases, we even approach the complex model's performance. We also conduct fu rther experiments to validate assertions and intuitively understand why our meth od works. Theoretically, we motivate our approach by showing that the weighted l oss minimized by simple models using our weighting upper bounds the loss of the complex model.

****

Modeling treatment events in disease progression

Guanyang Wang,Yumeng Zhang,Yong Deng,Xuxin Huang,Lukasz Kidzinski

Ability to quantify and predict progression of a disease is fundamental for sele cting an appropriate treatment. Many clinical metrics cannot be acquired frequen tly either because of their cost (e.g. MRI, gait analysis) or because they are i nconvenient or harmful to a patient (e.g. biopsy, x-ray). In such scenarios, in order to estimate individual trajectories of disease progression, it is advantag eous to leverage similarities between patients, i.e. the covariance of trajector ies, and find a latent representation of progression. Most of existing methods f or estimating trajectories do not account for events in-between observations, wh at dramatically decreases their adequacy for clinical practice. In this study, w e develop a machine learning framework named Coordinatewise-Soft-Impute (CSI) fo r analyzing disease progression from sparse observations in the presence of conf ounding events. CSI is guaranteed to converge to the global minimum of the corre sponding optimization problem. Experimental results also demonstrates the effect iveness of CSI using both simulated and real dataset.

****

DG-GAN: the GAN with the duality gap

Cheng Peng,Hao Wang,Xiao Wang,Zhouwang Yang

Generative Adversarial Networks (GANs) are powerful, but difficult to understand and train because GANs is a min-max problem. This paper understand GANs with d uality gap that comes from game theorem and show that duality gap can be a kind of metric to evolution the difference between the true data distribution and the distribution generated by generator with given condition. And train the network s using duality gap can get some better results. Furthermore, the paper calculat es the generalization bound of duality gap to estimate the help design the neura l networks and select the sample size.

****

Stochastic Gradient Descent with Biased but Consistent Gradient Estimators

Jie Chen,Ronny Luss

Stochastic gradient descent (SGD), which dates back to the 1950s, is one of the most popular and effective approaches for performing stochastic optimization. Re search on SGD resurged recently in machine learning for optimizing convex loss f unctions and training nonconvex deep neural networks. The theory assumes that on e can easily compute an unbiased gradient estimator, which is usually the case d ue to the sample average nature of empirical risk minimization. There exist, how ever, many scenarios (e.g., graphs) where an unbiased estimator may be as expens ive to compute as the full gradient because training examples are interconnected . Recently, Chen et al. (2018) proposed using a consistent gradient estimator as an economic alternative. Encouraged by empirical success, we show, in a general setting, that consistent estimators result in the same convergence behavior as do unbiased ones. Our analysis covers strongly convex, convex, and nonconvex obj ectives. We verify the results with illustrative experiments on synthetic and re al-world data. This work opens several new research directions, including the de velopment of more efficient SGD updates with consistent estimators and the desig n of efficient training algorithms for large-scale graphs.

****

One-way prototypical networks

Anna Kruspe

Few-shot models have become a popular topic of research in the past years. They offer the possibility to determine class belongings for unseen examples using ju

st a handful of examples for each class. Such models are trained on a wide range of classes and their respective examples, learning a decision metric in the process. Types of few-shot models include matching networks and prototypical networks. We show a new way of training prototypical few-shot models for just a single class. These models have the ability to predict the likelihood of an unseen query belonging to a group of examples without any given counterexamples. The difficulty here lies in the fact that no relative distance to other classes can be calculated
via softmax. We solve this problem by introducing a “null class” centered around zero, and enforcing centering with batch normalization. Trained on the commonly used Omniglot data set, we obtain a classification accuracy of .98 on the matched
test set, and of .8 on unmatched MNIST data. On the more complex MiniImageNet data set, test accuracy is .8. In addition, we propose a novel Gaussian layer for distance calculation in a prototypical network, which takes the support examples’ distribution rather than just their centroid into account. This extension shows promising results when a higher number of support examples is available.
****

Encoding word order in complex embeddings
Benyou Wang,Donghao Zhao,Christina Lioma,Qiuchi Li,Peng Zhang,Jakob Grue Simonsen
Sequential word order is important when processing text. Currently, neural networks (NNs) address this by modeling word position using position embeddings. The problem is that position embeddings capture the position of individual words, but not the ordered relationship (e.g., adjacency or precedence) between individual word positions. We present a novel and principled solution for modeling both the global absolute positions of words and their order relationships. Our solution generalizes word embeddings, previously defined as independent vectors, to continuous word functions over a variable (position). The benefit of continuous functions over variable positions is that word representations shift smoothly with increasing positions. Hence, word representations in different positions can correlate with each other in a continuous function. The general solution of these functions can be extended to complex-valued variants. We extend CNN, RNN and Transformer NNs to complex-valued versions to incorporate our complex embedding (we make all code available). Experiments on text classification, machine translation and language modeling show gains over both classical word embeddings and position-enriched word embeddings. To our knowledge, this is the first work in NLP to link imaginary numbers in complex-valued representations to concrete meanings (i.e., word order).
****

Functional vs. parametric equivalence of ReLU networks
Mary Phuong,Christoph H. Lampert
We address the following question: How redundant is the parameterisation of ReLU networks? Specifically, we consider transformations of the weight space which leave the function implemented by the network intact. Two such transformations are known for feed-forward architectures: permutation of neurons within a layer, and positive scaling of all incoming weights of a neuron coupled with inverse scaling of its outgoing weights. In this work, we show for architectures with non-increasing widths that permutation and scaling are in fact the only function-preserving weight transformations. For any eligible architecture we give an explicit construction of a neural network such that any other network that implements the same function can be obtained from the original one by the application of permutations and rescaling. The proof relies on a geometric understanding of boundaries between linear regions of ReLU networks, and we hope the developed mathematical tools are of independent interest.

****

A New Multi-input Model with the Attention Mechanism for Text Classification
Junhao Qiu,Ronghua Shi,Fangfang Li (the corresponding author),Jinjing Shi,Wangmin Liao

Recently, deep learning has made extraordinary achievements in text classificati
on. However, most of present models, especially convolutional neural network (CN
N), do not extract long-range associations, global representations, and hierarch
ical features well due to their relatively shallow and simple structures. This c
auses a negative effect on text classification. Moreover, we find that there are
 many express methods of texts. It is appropriate to design the multi-input mode
l to improve the classification effect. But most of models of text classificatio
n only use words or characters and do not use the multi-input model. Inspired by
 the above points and Densenet (Gao Huang, Zhuang Liu, Laurens Van Der Maaten, a
nd Kilian Q Weinberger. Densely connected convolutional networks. In Proceedings
 of the IEEE conference on computer vision and pattern recognition, pp. 4700–470
8, 2017.), we propose a new text classification model, which uses words, charact
ers, and labels as input. The model, which is a deep CNN with a novel attention
mechanism, can effectively leverage the input information and solve the above is
sues of the shallow model. We conduct experiments on six large text classificati
on datasets. Our model achieves the state of the art results on all datasets com
pared to multiple baseline models.
****

Multi-Dimensional Explanation of Reviews
Diego Antognini,Claudiu Musat,Boi Faltings
Neural models achieved considerable improvement for many natural language proces
sing tasks, but they offer little transparency, and interpretability comes at a
cost. In some domains, automated predictions without justifications have limited
 applicability. Recently, progress has been made regarding single-aspect sentime
nt analysis for reviews, where the ambiguity of a justification is minimal. In t
his context, a justification, or mask, consists of (long) word sequences from th
e input text, which suffice to make the prediction. Existing models cannot handl
e more than one aspect in one training and induce binary masks that might be amb
iguous. In our work, we propose a neural model for predicting multi-aspect senti
ments for reviews and generates a probabilistic multi-dimensional mask (one per
aspect) simultaneously, in an unsupervised and multi-task learning manner. Our e
valuation shows that on three datasets, in the beer and hotel domain, our model
outperforms strong baselines and generates masks that are: strong feature predic
tors, meaningful, and interpretable.
****

A Uniform Generalization Error Bound for Generative Adversarial Networks
Hao Chen,Zhanfeng Mo,Qingyi Gao,Zhouwang Yang,Xiao Wang
■This paper focuses on the theoretical investigation of unsupervised generalizat
ion theory of generative adversarial networks (GANs). We first formulate a more
reasonable definition of general error and generalization bounds for GANs. On to
p of that, we establish a bound for generalization error with a fixed generator
 in a general weight normalization context. Then, we obtain a width-independent
bound by applying $\ell_{p,q}$ and spectral norm weight normalization. To better
 understand the unsupervised model, GANs, we establish the generalization bound
, which uniformly holds with respect to the choice of generators. Hence, we can
explain how the complexity of discriminators and generators contribute to genera
lization error. For $\ell_{p,q}$ and spectral weight normalization, we provide e
xplicit guidance on how to design parameters to train robust generators. Our num
erical simulations also verify that our generalization bound is reasonable.
****

QGAN: Quantize Generative Adversarial Networks to Extreme low-bits
Peiqi Wang,Yu Ji,Xinfeng Xie,Yongqiang Lyu,Dongsheng Wang,Yuan Xie
The intensive computation and memory requirements of generative adversarial neur
al networks (GANs) hinder its real-world deployment on edge devices such as smar
tphones. Despite the success in model reduction of convolutional neural networks
 (CNNs), neural network quantization methods have not yet been studied on GANs,
which are mainly faced with the issues of both the effectiveness of quantization
 algorithms and the instability of training GAN models. In this paper, we start
with an extensive study on applying existing successful CNN quantization methods
 to quantize GAN models to extreme low bits. Our observation reveals that none o

f them generates samples with reasonable quality because of the underrepresentat ion of quantized weights in models, and the generator and discriminator networks show different sensitivities upon the quantization precision. Motivated by thes e observations, we develop a novel quantization method for GANs based on EM algo rithms, named as QGAN. We also propose a multi-precision algorithm to help find an appropriate quantization precision of GANs given image qualities requirements . Experiments on CIFAR-10 and CelebA show that QGAN can quantize weights in GANs to even 1-bit or 2-bit representations with results of quality comparable to or iginal models.
****

Contrastive Learning of Structured World Models
Thomas Kipf,Elise van der Pol,Max Welling
A structured understanding of our world in terms of objects, relations, and hier archies is an important component of human cognition. Learning such a structured world model from raw sensory data remains a challenge. As a step towards this g oal, we introduce Contrastively-trained Structured World Models (C-SWMs). C-SWMs utilize a contrastive approach for representation learning in environments with compositional structure. We structure each state embedding as a set of object r epresentations and their relations, modeled by a graph neural network. This allo ws objects to be discovered from raw pixel observations without direct supervisi on as part of the learning process. We evaluate C-SWMs on compositional environm ents involving multiple interacting objects that can be manipulated independentl y by an agent, simple Atari games, and a multi-object physics simulation. Our ex periments demonstrate that C-SWMs can overcome limitations of models based on pi xel reconstruction and outperform typical representatives of this model class in highly structured environments, while learning interpretable object-based repre sentations.
****

Disentangling Factors of Variations Using Few Labels
Francesco Locatello,Michael Tschannen,Stefan Bauer,Gunnar Rätsch,Bernhard Schölk opf,Olivier Bachem
Learning disentangled representations is considered a cornerstone problem in rep resentation learning. Recently, Locatello et al. (2019) demonstrated that unsupe rvised disentanglement learning without inductive biases is theoretically imposs ible and that existing inductive biases and unsupervised methods do not allow to consistently learn disentangled representations. However, in many practical set tings, one might have access to a limited amount of supervision, for example thr ough manual labeling of (some) factors of variation in a few training examples. In this paper, we investigate the impact of such supervision on state-of-the-art disentanglement methods and perform a large scale study, training over 52000 mo dels under well-defined and reproducible experimental conditions. We observe th at a small number of labeled examples (0.01--0.5% of the data set), with potenti ally imprecise and incomplete labels, is sufficient to perform model selection o n state-of-the-art unsupervised models. Further, we investigate the benefit of i ncorporating supervision into the training process. Overall, we empirically vali date that with little and imprecise supervision it is possible to reliably learn disentangled representations.
****

Detecting Out-of-Distribution Inputs to Deep Generative Models Using Typicality
Eric Nalisnick,Akihiro Matsukawa,Yee Whye Teh,Balaji Lakshminarayanan
Recent work has shown that deep generative models can assign higher likelihood t o out-of-distribution data sets than to their training data [Nalisnick et al., 2 019; Choi et al., 2019]. We posit that this phenomenon is caused by a mismatch between the model's typical set and its areas of high probability density. In-d istribution inputs should reside in the former but not necessarily in the latter , as previous work has presumed [Bishop, 1994]. To determine whether or not inp uts reside in the typical set, we propose a statistically principled, easy-to-im plement test using the empirical distribution of model likelihoods. The test is model agnostic and widely applicable, only requiring that the likelihood can be computed or closely approximated. We report experiments showing that our proce

dure can successfully detect the out-of-distribution sets in several of the chal lenging cases reported by Nalisnick et al. [2019].
****
EDUCE: Explaining model Decision through Unsupervised Concepts Extraction
Diane Bouchacourt,Ludovic Denoyer
Providing explanations along with predictions is crucial in some text processing tasks. Therefore, we propose a new self-interpretable model that performs outpu t prediction and simultaneously provides an explanation in terms of the presence of particular concepts in the input. To do so, our model's prediction relies so lely on a low-dimensional binary representation of the input, where each feature denotes the presence or absence of concepts. The presence of a concept is decid ed from an excerpt i.e. a small sequence of consecutive words in the text. Relev ant concepts for the prediction task at hand are automatically defined by our mo del, avoiding the need for concept-level annotations. To ease interpretability, we enforce that for each concept, the corresponding excerpts share similar seman tics and are differentiable from each others. We experimentally demonstrate the relevance of our approach on text classification and multi-sentiment analysis ta sks.
****
A critical analysis of self-supervision, or what we can learn from a single imag e
Asano YM.,Rupprecht C.,Vedaldi A.
We look critically at popular self-supervision techniques for learning deep conv olutional neural networks without manual labels. We show that three different an d representative methods, BiGAN, RotNet and DeepCluster, can learn the first few layers of a convolutional network from a single image as well as using millions of images and manual labels, provided that strong data augmentation is used. Ho wever, for deeper layers the gap with manual supervision cannot be closed even i f millions of unlabelled images are used for training.
We conclude that:
(1) the weights of the early layers of deep networks contain limited information about the statistics of natural images, that
(2) such low-level statistics can be learned through self-supervision just as we ll as through strong supervision, and that
(3) the low-level statistics can be captured via synthetic transformations inste ad of using a large image dataset.
****
Accelerating SGD with momentum for over-parameterized learning
Chaoyue Liu,Mikhail Belkin

Nesterov SGD is widely used for training modern neural networks and other machin e learning models. Yet, its advantages over SGD have not been theoretically clar ified. Indeed, as we show in this paper, both theoretically and empirically, Ne sterov SGD with any parameter selection does not in general provide acceleration over ordinary SGD. Furthermore, Nesterov SGD may diverge for step sizes that en sure convergence of ordinary SGD. This is in contrast to the classical results i n the deterministic setting, where the same step size ensures accelerated conver gence of the Nesterov's method over optimal gradient descent.

To address the non-acceleration issue, we introduce a compensation term to Nest erov SGD. The resulting algorithm, which we call MaSS, converges for same step sizes as SGD. We prove that MaSS obtains an accelerated convergence rates over SGD for any mini-batch size in the linear setting. For full batch, the converge nce rate of MaSS matches the well-known accelerated rate of the Nesterov's metho d.

We also analyze the practically important question of the dependence of the con vergence rate and optimal hyper-parameters on the mini-batch size, demonstratin g three distinct regimes: linear scaling, diminishing returns and saturation.

Experimental evaluation of MaSS for several standard architectures of deep networks, including ResNet and convolutional networks, shows improved performance over SGD, Nesterov SGD and Adam.
****

Discrete InfoMax Codes for Meta-Learning
Yoonho Lee,Wonjae Kim,Seungjin Choi
This paper analyzes how generalization works in meta-learning. Our core contribution is an information-theoretic generalization bound for meta-learning, which identifies the expressivity of the task-specific learner as the key factor that makes generalization to new datasets difficult. Taking inspiration from our bound, we present Discrete InfoMax Codes (DIMCO), a novel meta-learning model that trains a stochastic encoder to output discrete codes. Experiments show that DIMCO requires less memory and less time for similar performance to previous metric learning methods and that our method generalizes particularly well in a challenging small-data setting.
****

The Geometry of Sign Gradient Descent
Lukas Balles,Fabian Pedregosa,Nicolas Le Roux
Sign gradient descent has become popular in machine learning due to its favorable communication cost in distributed optimization and its good performance in neural network training. However, we currently do not have a good understanding of which geometrical properties of the objective function determine the relative speed of sign gradient descent compared to standard gradient descent. In this work, we frame sign gradient descent as steepest descent with respect to the maximum norm. We review the steepest descent framework and the related concept of smoothness with respect to arbitrary norms.
By studying the smoothness constant resulting from the L^∞-geometry, we isolate properties of the objective which favor sign gradient descent relative to gradient descent. In short, we find two requirements on its Hessian: (i) some degree of ``diagonal dominance'' and (ii) the maximal eigenvalue being much larger than the average eigenvalue. We also clarify the meaning of a certain separable smoothness assumption used in previous analyses of sign gradient descent.
Experiments verify the developed theory.
****

Learning Temporal Coherence via Self-Supervision for GAN-based Video Generation
Mengyu Chu,You Xie,Jonas Mayer,Laura Leal-Taixé,Nils Thürey
We focus on temporal self-supervision for GAN-based video generation tasks. While adversarial training successfully yields generative models for a variety of areas, temporal relationship in the generated data is much less explored. This is crucial for sequential generation tasks, e.g. video super-resolution and unpaired video translation. For the former, state-of-the-art methods often favor simpler norm losses such as L2 over adversarial training. However, their averaging nature easily leads to temporally smooth results with an undesirable lack of spatial detail. For unpaired video translation, existing approaches modify the generator networks to form spatio-temporal cycle consistencies. In contrast, we focus on improving the learning objectives and propose a temporally self-supervised algorithm. For both tasks, we show that temporal adversarial learning is key to achieving temporally coherent solutions without sacrificing spatial detail. We also propose a novel Ping-Pong loss to improve the long-term temporal consistency. It effectively prevents recurrent networks from accumulating artifacts temporally without depressing detailed features. We also propose a first set of metrics to quantitatively evaluate the accuracy as well as the perceptual quality of the temporal evolution. A series of user studies confirms the rankings computed with these metrics.
****

Interpretable Complex-Valued Neural Networks for Privacy Protection
Liyao Xiang,Hao Zhang,Haotian Ma,Yifan Zhang,Jie Ren,Quanshi Zhang
Previous studies have found that an adversary attacker can often infer unintended input information from intermediate-layer features. We study the possibility of preventing such adversarial inference, yet without too much accuracy degradati

on. We propose a generic method to revise the neural network to boost the challe nge of inferring input attributes from features, while maintaining highly accura te outputs. In particular, the method transforms real-valued features into compl ex-valued ones, in which the input is hidden in a randomized phase of the transf ormed features. The knowledge of the phase acts like a key, with which any party can easily recover the output from the processing result, but without which the party can neither recover the output nor distinguish the original input. Prelim inary experiments on various datasets and network structures have shown that our method significantly diminishes the adversary's ability in inferring about the input while largely preserves the resulting accuracy.
****

V-MPO: On-Policy Maximum a Posteriori Policy Optimization for Discrete and Conti nuous Control
H. Francis Song,Abbas Abdolmaleki,Jost Tobias Springenberg,Aidan Clark,Hubert So yer,Jack W. Rae,Seb Noury,Arun Ahuja,Siqi Liu,Dhruva Tirumala,Nicolas Heess,Dan Belov,Martin Riedmiller,Matthew M. Botvinick
Some of the most successful applications of deep reinforcement learning to chall enging domains in discrete and continuous control have used policy gradient meth ods in the on-policy setting. However, policy gradients can suffer from large va riance that may limit performance, and in practice require carefully tuned entro py regularization to prevent policy collapse. As an alternative to policy gradie nt algorithms, we introduce V-MPO, an on-policy adaptation of Maximum a Posterio ri Policy Optimization (MPO) that performs policy iteration based on a learned s tate-value function. We show that V-MPO surpasses previously reported scores for both the Atari-57 and DMLab-30 benchmark suites in the multi-task setting, and does so reliably without importance weighting, entropy regularization, or popula tion-based tuning of hyperparameters. On individual DMLab and Atari levels, the proposed algorithm can achieve scores that are substantially higher than has pre viously been reported. V-MPO is also applicable to problems with high-dimensiona l, continuous action spaces, which we demonstrate in the context of learning to control simulated humanoids with 22 degrees of freedom from full state observati ons and 56 degrees of freedom from pixel observations, as well as example OpenAI Gym tasks where V-MPO achieves substantially higher asymptotic scores than prev iously reported.
****

Improving Adversarial Robustness Requires Revisiting Misclassified Examples
Yisen Wang,Difan Zou,Jinfeng Yi,James Bailey,Xingjun Ma,Quanquan Gu
Deep neural networks (DNNs) are vulnerable to adversarial examples crafted by im perceptible perturbations. A range of defense techniques have been proposed to i mprove DNN robustness to adversarial examples, among which adversarial training has been demonstrated to be the most effective. Adversarial training is often fo rmulated as a min-max optimization problem, with the inner maximization for gene rating adversarial examples. However, there exists a simple, yet easily overlook ed fact that adversarial examples are only defined on correctly classified (natu ral) examples, but inevitably, some (natural) examples will be misclassified dur ing training. In this paper, we investigate the distinctive influence of misclas sified and correctly classified examples on the final robustness of adversarial training. Specifically, we find that misclassified examples indeed have a signif icant impact on the final robustness. More surprisingly, we find that different maximization techniques on misclassified examples may have a negligible influenc e on the final robustness, while different minimization techniques are crucial. Motivated by the above discovery, we propose a new defense algorithm called {\em Misclassification Aware adveRsarial Training} (MART), which explicitly differen tiates the misclassified and correctly classified examples during the training. We also propose a semi-supervised extension of MART, which can leverage the unla beled data to further improve the robustness. Experimental results show that MAR T and its variant could significantly improve the state-of-the-art adversarial r obustness.
****

InfoCNF: Efficient Conditional Continuous Normalizing Flow Using Adaptive Solver

s
Tan M. Nguyen,Animesh Garg,Richard G. Baraniuk,Anima Anandkumar
Continuous Normalizing Flows (CNFs) have emerged as promising deep generative models for a wide range of tasks thanks to their invertibility and exact likelihood estimation. However, conditioning CNFs on signals of interest for conditional image generation and downstream predictive tasks is inefficient due to the high-dimensional latent code generated by the model, which needs to be of the same size as the input data. In this paper, we propose InfoCNF, an efficient conditional CNF that partitions the latent space into a class-specific supervised code and an unsupervised code that shared among all classes for efficient use of labeled information. Since the partitioning strategy (slightly) increases the number of function evaluations (NFEs), InfoCNF also employs gating networks to learn the error tolerances of its ordinary differential equation (ODE) solvers for better speed and performance. We show empirically that InfoCNF improves the test accuracy over the baseline while yielding comparable likelihood scores and reducing the NFEs on CIFAR10. Furthermore, applying the same partitioning strategy in InfoCNF on time-series data helps improve extrapolation performance.
****

Mirror Descent View For Neural Network Quantization
Thalaiyasingam Ajanthan,Kartik Gupta,Philip H. S. Torr,Richard Hartley,Puneet K. Dokania
Quantizing large Neural Networks (NN) while maintaining the performance is highly desirable for resource-limited devices due to reduced memory and time complexity. NN quantization is usually formulated as a constrained optimization problem and optimized via a modified version of gradient descent. In this work, by interpreting the continuous parameters (unconstrained) as the dual of the quantized ones, we introduce a Mirror Descent (MD) framework (Bubeck (2015)) for NN quantization. Specifically, we provide conditions on the projections (i.e., mapping from continuous to quantized ones) which would enable us to derive valid mirror maps and in turn the respective MD updates. Furthermore, we discuss a numerically stable implementation of MD by storing an additional set of auxiliary dual variables (continuous). This update is strikingly analogous to the popular Straight Through Estimator (STE) based method which is typically viewed as a “trick” to avoid vanishing gradients issue but here we show that it is an implementation method for MD for certain projections. Our experiments on standard classification datasets (CIFAR-10/100, TinyImageNet) with convolutional and residual architectures show that our MD variants obtain fully-quantized networks with accuracies very close to the floating-point networks.
****

Hierarchical Disentangle Network for Object Representation Learning
Shishi Qiao,Ruiping Wang,Shiguang Shan,Xilin Chen
An object can be described as the combination of primary visual attributes. Disentangling such underlying primitives is the long objective of representation learning. It is observed that categories have the natural multi-granularity or hierarchical characteristics, i.e. any two objects can share some common primitives in a particular category granularity while they may possess their unique ones in another granularity. However, previous works usually operate in a flat manner (i.e. in a particular granularity) to disentangle the representations of objects. Though they may obtain the primitives to constitute objects as the categories in that granularity, their results are obviously not efficient and complete. In this paper, we propose the hierarchical disentangle network (HDN) to exploit the rich hierarchical characteristics among categories to divide the disentangling process in a coarse-to-fine manner, such that each level only focuses on learning the specific representations in its granularity and finally the common and unique representations in all granularities jointly constitute the raw object. Specifically, HDN is designed based on an encoder-decoder architecture. To simultaneously ensure the disentanglement and interpretability of the encoded representations, a novel hierarchical generative adversarial network (GAN) is elaborately designed. Quantitative and qualitative evaluations on four object datasets validate the effectiveness of our method.

****
Deep Multiple Instance Learning with Gaussian Weighting
Basura Fernando,Hakan Bilen
In this paper we present a deep Multiple Instance Learning (MIL) method that can be trained end-to-end to perform classification from weak supervision. Our MIL method is implemented as a two stream neural network, specialized in tasks of instance classification and weighting. Our instance weighting stream makes use of Gaussian radial basis function to normalize the instance weights by comparing instances locally within the bag and globally across bags. The final classification score of the bag is an aggregate of all instance classification scores. The instance representation is shared by both instance classification and weighting streams. The Gaussian instance weighting allows us to regularize the representation learning of instances such that all positive instances to be closer to each other w.r.t. the instance weighting function. We evaluate our method on five standard MIL datasets and show that our method outperforms other MIL methods. We also evaluate our model on two datasets where all models are trained end-to-end. Our method obtain better bag-classification and instance classification results on these datasets. We conduct extensive experiments to investigate the robustness of the proposed model and obtain interesting insights.
****
Zeno++: Robust Fully Asynchronous SGD
Cong Xie,Oluwasanmi Koyejo,Indranil Gupta
We propose Zeno++, a new robust asynchronous Stochastic Gradient Descent~(SGD) procedure which tolerates Byzantine failures of the workers. In contrast to previous work, Zeno++ removes some unrealistic restrictions on worker-server communications, allowing for fully asynchronous updates from anonymous workers, arbitrarily stale worker updates, and the possibility of an unbounded number of Byzantine workers. The key idea is to estimate the descent of the loss value after the candidate gradient is applied, where large descent values indicate that the update results in optimization progress. We prove the convergence of Zeno++ for non-convex problems under Byzantine failures. Experimental results show that Zeno++ outperforms existing approaches.
****
DivideMix: Learning with Noisy Labels as Semi-supervised Learning
Junnan Li,Richard Socher,Steven C.H. Hoi
Deep neural networks are known to be annotation-hungry. Numerous efforts have been devoted to reducing the annotation cost when learning with deep networks. Two prominent directions include learning with noisy labels and semi-supervised learning by exploiting unlabeled data. In this work, we propose DivideMix, a novel framework for learning with noisy labels by leveraging semi-supervised learning techniques. In particular, DivideMix models the per-sample loss distribution with a mixture model to dynamically divide the training data into a labeled set with clean samples and an unlabeled set with noisy samples, and trains the model on both the labeled and unlabeled data in a semi-supervised manner. To avoid confirmation bias, we simultaneously train two diverged networks where each network uses the dataset division from the other network. During the semi-supervised training phase, we improve the MixMatch strategy by performing label co-refinement and label co-guessing on labeled and unlabeled samples, respectively. Experiments on multiple benchmark datasets demonstrate substantial improvements over state-of-the-art methods. Code is available at https://github.com/LiJunnan1992/DivideMix .
****
Extreme Value k-means Clustering
Sixiao Zheng,Yanxi Hou,Yanwei Fu,Jianfeng Feng
Clustering is the central task in unsupervised learning and data mining. k-means is one of the most widely used clustering algorithms. Unfortunately, it is generally non-trivial to extend k-means to cluster data points beyond Gaussian distribution, particularly, the clusters with non-convex shapes (Beliakov & King, 2006). To this end, we, for the first time, introduce Extreme Value Theory (EVT) to improve the clustering ability of k-means. Particularly, the Euclidean space wa

s transformed into a novel probability space denoted as extreme value space by E
VT. We thus propose a novel algorithm called Extreme Value k-means (EV k-means),
 including GEV k-means and GPD k-means. In addition, we also introduce the trick
s to accelerate Euclidean distance computation in improving the computational ef
ficiency of classical k-means. Furthermore, our EV k-means is extended to an onl
ine version, i.e., online Extreme Value k-means, in utilizing the Mini Batch k-m
eans to cluster streaming data. Extensive experiments are conducted to validate
our EV k-means and online EV k-means on synthetic datasets and real datasets. Ex
perimental results show that our algorithms significantly outperform competitors
 in most cases.
****

Adaptive network sparsification with dependent variational beta-Bernoulli dropou
t

Juho Lee,Saehoon Kim,Jaehong Yoon,Hae Beom Lee,Eunho Yang,Sung Ju Hwang

While variational dropout approaches have been shown to be effective for network
 sparsification, they are still suboptimal in the sense that they set the dropou
t rate for each neuron without consideration of the input data. With such input
independent dropout, each neuron is evolved to be generic across inputs, which m
akes it difficult to sparsify networks without accuracy loss. To overcome this l
imitation, we propose adaptive variational dropout whose probabilities are drawn
 from sparsity inducing beta-Bernoulli prior. It allows each neuron to be evolve
d either to be generic or specific for certain inputs, or dropped altogether. Su
ch input-adaptive sparsity- inducing dropout allows the resulting network to tol
erate larger degree of sparsity without losing its expressive power by removing
redundancies among features. We validate our dependent variational beta-Bernoull
i dropout on multiple public datasets, on which it obtains significantly more co
mpact networks than baseline methods, with consistent accuracy improvements over
 the base networks.
****

Data-dependent Gaussian Prior Objective for Language Generation

Zuchao Li,Rui Wang,Kehai Chen,Masso Utiyama,Eiichiro Sumita,Zhuosheng Zhang,Hai
Zhao

For typical sequence prediction problems such as language generation, maximum li
kelihood estimation (MLE) has commonly been adopted as it encourages the predict
ed sequence most consistent with the ground-truth sequence to have the highest p
robability of occurring. However, MLE focuses on once-to-all matching between th
e predicted sequence and gold-standard, consequently treating all incorrect pred
ictions as being equally incorrect. We refer to this drawback as {\it negative d
iversity ignorance} in this paper. Treating all incorrect predictions as equal u
nfairly downplays the nuance of these sequences' detailed token-wise structure.
To counteract this, we augment the MLE loss by introducing an extra Kullback--Le
ibler divergence term derived by comparing a data-dependent Gaussian prior and t
he detailed training prediction. The proposed data-dependent Gaussian prior obje
ctive (D2GPo) is defined over a prior topological order of tokens and is poles a
part from the data-independent Gaussian prior (L2 regularization) commonly adopt
ed in smoothing the training of MLE. Experimental results show that the proposed
 method makes effective use of a more detailed prior in the data and has improve
d performance in typical language generation tasks, including supervised and uns
upervised machine translation, text summarization, storytelling, and image capti
oning.

****

Learning Representations in Reinforcement Learning: an Information Bottleneck Ap
proach

Yingjun Pei,Xinwen Hou

The information bottleneck principle is an elegant and useful approach to repres
entation learning. In this paper, we investigate the problem of representation l
earning in the context of reinforcement learning using the information bottlenec
k framework, aiming at improving the sample efficiency of the learning algorithm
s.We analytically derive the optimal conditional distribution of the representat

ion, and provide a variational lower bound. Then, we maximize this lower bound w
ith the Stein variational (SV) gradient method.
We incorporate this framework in the advantageous actor critic algorithm (A2C) a
nd the proximal policy optimization algorithm (PPO). Our experimental results sh
ow that our framework can improve the sample efficiency of vanilla A2C and PPO s
ignificantly. Finally, we study the information-bottleneck (IB) perspective in d
eep RL with the algorithm called mutual information neural estimation(MINE).
We experimentally verify that the information extraction-compression process als
o exists in deep RL and our framework is capable of accelerating this process. W
e also analyze the relationship between MINE and our method, through this relati
onship, we theoretically derive an algorithm to optimize our IB framework withou
t constructing the lower bound.
****

LSTOD: Latent Spatial-Temporal Origin-Destination prediction model and its appli
cations in ride-sharing platforms
Fan Zhou,Haibo Zhou,Hongtu Zhu
Origin-Destination (OD) flow data is an important instrument in transportation s
tudies. Precise prediction of customer demands from each original location to a
destination given a series of previous snapshots helps ride-sharing platforms t
o better understand their market mechanism. However, most existing prediction me
thods ignore the network structure of OD flow data and fail to utilize the topol
ogical dependencies among related OD pairs. In this paper, we propose a latent s
patial-temporal origin-destination (LSTOD) model, with a novel convolutional neu
ral network (CNN) filter to learn the spatial features of OD pairs from a graph
perspective and an attention structure to capture their long-term periodicity. E
xperiments on a real customer request dataset with available OD information from
 a ride-sharing platform demonstrate the advantage of LSTOD in achieving at leas
t 6.5% improvement in prediction accuracy over the second best model.
****

Ecological Reinforcement Learning
John D. Co-Reyes,Suvansh Sanjeev,Glen Berseth,Abhishek Gupta,Sergey Levine
Reinforcement learning algorithms have been shown to effectively learn tasks in
a variety of static, deterministic, and simplistic environments, but their appl
ication to environments which are characteristic of dynamic lifelong settings en
countered in the real world has been limited. Understanding the impact of specif
ic environmental properties on the learning dynamics of reinforcement learning a
lgorithms is important as we want to align the environments in which we develop
our algorithms with the real world, and this is strongly coupled with the type o
f intelligence which can be learned. In this work, we study what we refer to as
ecological reinforcement learning: the interaction between properties of the env
ironment and the reinforcement learning agent. To this end, we introduce environ
ments with characteristics that we argue better reflect natural environments: no
n-episodic learning, uninformative ``fundamental drive'' reward signals, and nat
ural dynamics that cause the environment to change even when the agent fails to
take intelligent actions. We show these factors can have a profound effect on th
e learning progress of reinforcement learning algorithms. Surprisingly, we find
that these seemingly more challenging learning conditions can often make reinfor
cement learning agents learn more effectively. Through this study, we hope to sh
ift the focus of the community towards learning in realistic, natural environmen
ts with dynamic elements.
****

Towards Understanding the Regularization of Adversarial Robustness on Neural Net
works
Yuxin Wen,Shuai Li,Kui Jia
  The problem of adversarial examples has shown that modern Neural Network (NN)
models could be rather fragile. Among the most promising techniques to solve the
 problem, one is to require the model to be {\it ϵ-adversarially robust
} (AR); that is, to require the model not to change predicted labels when any gi
ven input examples are perturbed within a certain range. However, it is widely o
bserved that such methods would lead to standard performance degradation, i.e.,

the degradation on natural examples. In this work, we study the degradation thr
ough the regularization perspective. We identify quantities from generalization
analysis of NNs; with the identified quantities we empirically find that AR is
achieved by regularizing/biasing NNs towards less confident solutions by making
the changes in the feature space (induced by changes in the instance space) of m
ost layers smoother uniformly in all directions; so to a certain extent, it prev
ents sudden change in prediction w.r.t. perturbations. However, the end result
of such smoothing concentrates samples around decision boundaries, resulting in
less confident solutions, and leads to worse standard performance. Our studies
suggest that one might consider ways that build AR into NNs in a gentler way to
avoid the problematic regularization.

****

MaskConvNet: Training Efficient ConvNets from Scratch via Budget-constrained Fil
ter Pruning

Raden Mu'az Mun'im,Jie Lin,Vijay Chandrasekhar,Koichi Shinoda

In this paper, we propose a framework, called MaskConvNet, for ConvNets filter p
runing. MaskConvNet provides elegant support for training budget-aware pruned ne
tworks from scratch, by adding a simple mask module to a ConvNet architecture. M
askConvNet enjoys several advantages - (1) Flexible, the mask module can be inte
grated with any ConvNets in a plug-and-play manner. (2) Simple, the mask module
is implemented by a hard Sigmoid function with a small number of trainable mask
variables, adding negligible memory and computational overheads to the networks
during training. (3) Effective, it is able to achieve competitive pruning rate w
hile maintaining comparable accuracy with the baseline ConvNets without pruning,
regardless of the datasets and ConvNet architectures used. (4) Fast, it is obse
rved that the number of training epochs required by MaskConvNet is close to trai
ning a baseline without pruning. (5) Budget-aware, with a sparsity budget on tar
get metric (e.g. model size and FLOP), MaskConvNet is able to train in a way tha
t the optimizer can adaptively sparsify the network and automatically maintain s
parsity level, till the pruned network produces good accuracy and fulfill the bu
dget constraint simultaneously. Results on CIFAR-10 and ImageNet with several Co
nvNet architectures show that MaskConvNet works competitively well compared to p
revious pruning methods, with budget-constraint well respected. Code is availabl
e at https://www.dropbox.com/s/c4zi3n7h1bexl12/maskconv-iclr-code.zip?dl=0. We h
ope MaskConvNet, as a simple and general pruning framework, can address the gaps
in existing literate and advance future studies to push the boundaries of neura
l network pruning.

****

Scale-Equivariant Neural Networks with Decomposed Convolutional Filters

Wei Zhu,Qiang Qiu,Robert Calderbank,Guillermo Sapiro,Xiuyuan Cheng

Encoding the input scale information explicitly into the representation learned
by a convolutional neural network (CNN) is beneficial for many vision tasks espe
cially when dealing with multiscale input signals. We study, in this paper, a sc
ale-equivariant CNN architecture with joint convolutions across the space and th
e scaling group, which is shown to be both sufficient and necessary to achieve s
cale-equivariant representations. To reduce the model complexity and computation
al burden, we decompose the convolutional filters under two pre-fixed separable
bases and truncate the expansion to low-frequency components. A further benefit
of the truncated filter expansion is the improved deformation robustness of the
equivariant representation. Numerical experiments demonstrate that the proposed
scale-equivariant neural network with decomposed convolutional filters (ScDCFNet
) achieves significantly improved performance in multiscale image classification
and better interpretability than regular CNNs at a reduced model size.

****

A novel Bayesian estimation-based word embedding model for sentiment analysis

Jingyao Tang,Yun Xue,Ziwen Wang,Haoliang Zhao

The word embedding models have achieved state-of-the-art results in a variety of
natural language processing tasks. Whereas, current word embedding models mainl
y focus on the rich semantic meanings while are challenged by capturing the sent

iment information. For this reason, we propose a novel sentiment word embedding model. In line with the working principle, the parameter estimating method is highlighted. On the task of semantic and sentiment embeddings, the parameters in the proposed model are determined by using both the maximum likelihood estimation and the Bayesian estimation. Experimental results show the proposed model significantly outperforms the baseline methods in sentiment analysis for low-frequency words and sentences. Besides, it is also effective in conventional semantic and sentiment analysis tasks.
****

Attacking Lifelong Learning Models with Gradient Reversion

Yunhui Guo,Mingrui Liu,Yandong Li,Liqiang Wang,Tianbao Yang,Tajana Rosing

Lifelong learning aims at avoiding the catastrophic forgetting problem of traditional supervised learning models. Episodic memory based lifelong learning methods such as A-GEM (Chaudhry et al., 2018b) are shown to achieve the state-of-the-art results across the benchmarks. In A-GEM, a small episodic memory is utilized to store a random subset of the examples from previous tasks. While the model is trained on a new task, a reference gradient is computed on the episodic memory to guide the direction of the current update. While A-GEM has strong continual learning ability, it is not clear that if it can retain the performance in the presence of adversarial attacks. In this paper, we examine the robustness ofA-GEM against adversarial attacks to the examples in the episodic memory. We evaluate the effectiveness of traditional attack methods such as FGSM and PGD.The results show that A-GEM still possesses strong continual learning ability in the presence of adversarial examples in the memory and simple defense techniques such as label smoothing can further alleviate the adversarial effects. We presume that traditional attack methods are specially designed for standard supervised learning models rather than lifelong learning models. we therefore propose a principled way for attacking A-GEM called gradient reversion (GREV) which is shown to be more effective. Our results indicate that future lifelong learning research should bear adversarial attacks in mind to develop more robust lifelong learning algorithms.
****

Learning with Long-term Remembering: Following the Lead of Mixed Stochastic Gradient

Yunhui Guo,Mingrui Liu,Tianbao Yang,Tajana Rosing

Current deep neural networks can achieve remarkable performance on a single task. However, when the deep neural network is continually trained on a sequence of tasks, it seems to gradually forget the previous learned knowledge. This phenomenon is referred to as catastrophic forgetting and motivates the field called lifelong learning. The central question in lifelong learning is how to enable deep neural networks to maintain performance on old tasks while learning a new task. In this paper, we introduce a novel and effective lifelong learning algorithm, called MixEd stochastic GrAdient (MEGA), which allows deep neural networks to acquire the ability of retaining performance on old tasks while learning new tasks. MEGA modulates the balance between old tasks and the new task by integrating the current gradient with the gradient computed on a small reference episodic memory. Extensive experimental results show that the proposed MEGA algorithm significantly advances the state-of-the-art on all four commonly used life-long learning benchmarks, reducing the error by up to 18%.
****

Fooling Detection Alone is Not Enough: Adversarial Attack against Multiple Object Tracking

Yunhan Jia,Yantao Lu,Junjie Shen,Qi Alfred Chen,Hao Chen,Zhenyu Zhong,Tao Wei

Recent work in adversarial machine learning started to focus on the visual perception in autonomous driving and studied Adversarial Examples (AEs) for object detection models. However, in such visual perception pipeline the detected objects must also be tracked, in a process called Multiple Object Tracking (MOT), to build the moving trajectories of surrounding obstacles. Since MOT is designed to be robust against errors in object detection, it poses a general challenge to existing attack techniques that blindly target objection detection: we find that a

success rate of over 98% is needed for them to actually affect the tracking resu lts, a requirement that no existing attack technique can satisfy. In this paper, we are the first to study adversarial machine learning attacks against the comp lete visual perception pipeline in autonomous driving, and discover a novel atta ck technique, tracker hijacking, that can effectively fool MOT using AEs on obje ct detection. Using our technique, successful AEs on as few as one single frame can move an existing object in to or out of the headway of an autonomous vehicle to cause potential safety hazards. We perform evaluation using the Berkeley Dee p Drive dataset and find that on average when 3 frames are attacked, our attack can have a nearly 100% success rate while attacks that blindly target object det ection only have up to 25%.

****

Towards A Unified Min-Max Framework for Adversarial Exploration and Robustness
Jingkang Wang,Tianyun Zhang,Sijia Liu,Pin-Yu Chen,Jiacen Xu,Makan Fardad,Bo Li
The worst-case training principle that minimizes the maximal adversarial loss, a lso known as adversarial training (AT), has shown to be a state-of-the-art appro ach for enhancing adversarial robustness against norm-ball bounded input perturb ations. Nonetheless, min-max optimization beyond the purpose of AT has not been rigorously explored in the research of adversarial attack and defense. In partic ular, given a set of risk sources (domains), minimizing the maximal loss induced from the domain set can be reformulated as a general min-max problem that is di fferent from AT. Examples of this general formulation include attacking model en sembles, devising universal perturbation under multiple inputs or data transform ations, and generalized AT over different types of attack models. We show that t hese problems can be solved under a unified and theoretically principled min-max optimization framework. We also show that the self-adjusted domain weights lea rned from our method provides a means to explain the difficulty level of attack and defense over multiple domains. Extensive experiments show that our approach leads to substantial performance improvement over the conventional averaging str ategy.

****

Domain-Agnostic Few-Shot Classification by Learning Disparate Modulators
Yongseok Choi,Junyoung Park,Subin Yi,Dong-Yeon Cho
Although few-shot learning research has advanced rapidly with the help of meta-l earning, its practical usefulness is still limited because most of the researche s assumed that all meta-training and meta-testing examples came from a single do main. We propose a simple but effective way for few-shot classification in which a task distribution spans multiple domains including previously unseen ones dur ing meta-training.

The key idea is to build a pool of embedding models which have their own metric spaces and to learn to select the best one for a particular task through multi-d omain meta-learning. This simplifies task-specific adaptation over a complex tas k distribution as a simple selection problem rather than modifying the model wit h a number of parameters at meta-testing time. Inspired by common multi-task lea rning techniques, we let all models in the pool share a base network and add a s eparate modulator to each model to refine the base network in its own way. This architecture allows the pool to maintain representational diversity and each mod el to have domain-invariant representation as well.

Experiments show that our selection scheme outperforms other few-shot classifica tion algorithms when target tasks could come from many different domains. They a lso reveal that aggregating outputs from all constituent models is effective for tasks from unseen domains showing the effectiveness of our framework.

****

Watch, Try, Learn: Meta-Learning from Demonstrations and Rewards
Allan Zhou,Eric Jang,Daniel Kappler,Alex Herzog,Mohi Khansari,Paul Wohlhart,Yunf ei Bai,Mrinal Kalakrishnan,Sergey Levine,Chelsea Finn
Imitation learning allows agents to learn complex behaviors from demonstrations. However, learning a complex vision-based task may require an impractical number of demonstrations. Meta-imitation learning is a promising approach towards enab ling agents to learn a new task from one or a few demonstrations by leveraging e

xperience from learning similar tasks. In the presence of task ambiguity or unob
served dynamics, demonstrations alone may not provide enough information; an age
nt must also try the task to successfully infer a policy. In this work, we propo
se a method that can learn to learn from both demonstrations and trial-and-error
 experience with sparse reward feedback. In comparison to meta-imitation, this a
pproach enables the agent to effectively and efficiently improve itself autonomo
usly beyond the demonstration data. In comparison to meta-reinforcement learning
, we can scale to substantially broader distributions of tasks, as the demonstra
tion reduces the burden of exploration. Our experiments show that our method sig
nificantly outperforms prior approaches on a set of challenging, vision-based co
ntrol tasks.
****

Logic and the 2-Simplicial Transformer
James Clift,Dmitry Doryn,Daniel Murfet,James Wallbridge
We introduce the 2-simplicial Transformer, an extension of the Transformer which
 includes a form of higher-dimensional attention generalising the dot-product at
tention, and uses this attention to update entity representations with tensor pr
oducts of value vectors. We show that this architecture is a useful inductive bi
as for logical reasoning in the context of deep reinforcement learning.

****

Reinforcement Learning with Chromatic Networks
Xingyou Song,Krzysztof Choromanski,Jack Parker-Holder,Yunhao Tang,Wenbo Gao,Aldo
 Pacchiano,Tamas Sarlos,Deepali Jain,Yuxiang Yang
We present a neural architecture search algorithm to construct compact reinforce
ment learning (RL) policies, by combining ENAS and ES in a highly scalable and i
ntuitive way. By defining the combinatorial search space of NAS to be the set of
 different edge-partitionings (colorings) into same-weight classes, we represent
 compact architectures via efficient learned edge-partitionings. For several RL
tasks, we manage to learn colorings translating to effective policies parameteri
zed by as few as 17 weight parameters, providing >90 % compression over vanilla
policies and 6x compression over state-of-the-art compact policies based on Toep
litz matrices, while still maintaining good reward. We believe that our work is
one of the first attempts to propose a rigorous approach to training structured
neural network architectures for RL problems that are of interest especially in
mobile robotics with limited storage and computational resources.
****

AE-OT: A NEW GENERATIVE MODEL BASED ON EXTENDED SEMI-DISCRETE OPTIMAL TRANSPORT
Dongsheng An,Yang Guo,Na Lei,Zhongxuan Luo,Shing-Tung Yau,Xianfeng Gu
Generative adversarial networks (GANs) have attracted huge attention due to
its capability to generate visual realistic images. However, most of the existin
g
models suffer from the mode collapse or mode mixture problems. In this work, we
give a theoretic explanation of the both problems by Figalli’s regularity theory
 of
optimal transportation maps. Basically, the generator compute the transportation
maps between the white noise distributions and the data distributions, which are
in general discontinuous. However, DNNs can only represent continuous maps.
This intrinsic conflict induces mode collapse and mode mixture. In order to
tackle the both problems, we explicitly separate the manifold embedding and the
optimal transportation; the first part is carried out using an autoencoder to ma
p the
images onto the latent space; the second part is accomplished using a GPU-based
convex optimization to find the discontinuous transportation maps. Composing the
extended OT map and the decoder, we can finally generate new images from the
white noise. This AE-OT model avoids representing discontinuous maps by DNNs,
therefore effectively prevents mode collapse and mode mixture.
****

Deep Mining: Detecting Anomalous Patterns in Neural Network Activations with Sub
set Scanning

Skyler Speakman,Celia Cintas,Victor Akinwande,Srihari Sridharan,Edward McFowland III

This work views neural networks as data generating systems and applies anomalous pattern detection techniques on that data in order to detect when a network is processing a group of anomalous inputs. Detecting anomalies is a critical component for multiple machine learning problems including detecting the presence of adversarial noise added to inputs. More broadly, this work is a step towards giving neural networks the ability to detect groups of out-of-distribution samples.

This work introduces ``Subset Scanning methods from the anomalous pattern detection domain to the task of detecting anomalous inputs to neural networks. Subset Scanning allows us to answer the question: "``Which subset of inputs have larger-than-expected activations at which subset of nodes?" Framing the adversarial detection problem this way allows us to identify systematic patterns in the activation space that span multiple adversarially noised images. Such images are ``"weird together". Leveraging this common anomalous pattern, we show increased detection power as the proportion of noised images increases in a test set. Detection power and accuracy results are provided for targeted adversarial noise added to CIFAR-10 images on a 20-layer ResNet using the Basic Iterative Method attack.

****

A Data-Efficient Mutual Information Neural Estimator for Statistical Dependency Testing

Xiao Lin,Indranil Sur,Samuel A. Nastase,Uri Hasson,Ajay Divakaran,Mohamed R. Amer

Measuring Mutual Information (MI) between high-dimensional, continuous, random variables from observed samples has wide theoretical and practical applications. Recent works have developed accurate MI estimators through provably low-bias approximations and tight variational lower bounds assuming abundant supply of samples, but require an unrealistic number of samples to guarantee statistical significance of the estimation. In this work, we focus on improving data efficiency and propose a Data-Efficient MINE Estimator (DEMINE) that can provide a tight lower confident interval of MI under limited data, through adding cross-validation to the MINE lower bound (Belghazi et al., 2018). Hyperparameter search is employed and a novel meta-learning approach with task augmentation is developed to increase robustness to hyperparamters, reduce overfitting and improve accuracy. With improved data-efficiency, our DEMINE estimator enables statistical testing of dependency at practical dataset sizes. We demonstrate the effectiveness of DEMINE on synthetic benchmarks and a real world fMRI dataset, with application of inter-subject correlation analysis.

****

Enhancing Adversarial Defense by k-Winners-Take-All

Chang Xiao,Peilin Zhong,Changxi Zheng

We propose a simple change to existing neural network structures for better defending against gradient-based adversarial attacks. Instead of using popular activation functions (such as ReLU), we advocate the use of k-Winners-Take-All (k-WTA) activation, a C0 discontinuous function that purposely invalidates the neural network model’s gradient at densely distributed input data points. The proposed k-WTA activation can be readily used in nearly all existing networks and training methods with no significant overhead. Our proposal is theoretically rationalized. We analyze why the discontinuities in k-WTA networks can largely prevent gradient-based search of adversarial examples and why they at the same time remain innocuous to the network training. This understanding is also empirically backed. We test k-WTA activation on various network structures optimized by a training method, be it adversarial training or not. In all cases, the robustness of k-WTA networks outperforms that of traditional networks under white-box attacks.

****

Thwarting finite difference adversarial attacks with output randomization

Haidar Khan,Dan Park,Azer Khan,Bülent Yener

Adversarial input poses a critical problem to deep neural networks (DNN). This problem is more severe in the "black box" setting where an adversary only needs

to repeatedly query a DNN to estimate the gradients required to create adversarial examples. Current defense techniques against attacks in this setting are not effective. Thus, in this paper, we present a novel defense technique based on randomization applied to a DNN's output layer. While effective as a defense technique, this approach introduces a trade off between accuracy and robustness. We show that for certain types of randomization, we can bound the probability of introducing errors by carefully setting distributional parameters. For the particular case of finite difference black box attacks, we quantify the error introduced by the defense in the finite difference estimate of the gradient. Lastly, we show empirically that the defense can thwart three adaptive black box adversarial attack algorithms.
****

Exploration in Reinforcement Learning with Deep Covering Options
Yuu Jinnai,Jee Won Park,Marlos C. Machado,George Konidaris
While many option discovery methods have been proposed to accelerate exploration in reinforcement learning, they are often heuristic. Recently, covering options was proposed to discover a set of options that provably reduce the upper bound of the environment's cover time, a measure of the difficulty of exploration. Covering options are computed using the eigenvectors of the graph Laplacian, but they are constrained to tabular tasks and are not applicable to tasks with large or continuous state-spaces.
We introduce deep covering options, an online method that extends covering options to large state spaces, automatically discovering task-agnostic options that encourage exploration. We evaluate our method in several challenging sparse-reward domains and we show that our approach identifies less explored regions of the state-space and successfully generates options to visit these regions, substantially improving both the exploration and the total accumulated reward.
****

Towards Controllable and Interpretable Face Completion via Structure-Aware and Frequency-Oriented Attentive GANs
Zeyuan Chen,Shaoliang Nie,Tianfu Wu,Christopher G. Healey
Face completion is a challenging conditional image synthesis task. This paper proposes controllable and interpretable high-resolution and fast face completion by learning generative adversarial networks (GANs) progressively from low resolution to high resolution. We present structure-aware and frequency-oriented attentive GANs. The proposed structure-aware component leverages off-the-shelf facial landmark detectors and proposes a simple yet effective method of integrating the detected landmarks in generative learning. It facilitates facial expression transfer together with facial attributes control, and helps regularize the structural consistency in progressive training. The proposed frequency-oriented attentive module (FOAM) encourages GANs to attend to only finer details in the coarse-to-fine progressive training, thus enabling progressive attention to face structures. The learned FOAMs show a strong pattern of switching its attention from low-frequency to high-frequency signals. In experiments, the proposed method is tested on the CelebA-HQ benchmark. Experiment results show that our approach outperforms state-of-the-art face completion methods. The proposed method is also fast with mean inference time of 0.54 seconds for images at 1024x1024 resolution (using a Titan Xp GPU).
****

Learning Disentangled Representations for CounterFactual Regression
Negar Hassanpour,Russell Greiner
We consider the challenge of estimating treatment effects from observational data; and point out that, in general, only some factors based on the observed covariates X contribute to selection of the treatment T, and only some to determining the outcomes Y. We model this by considering three underlying sources of {X, T, Y} and show that explicitly modeling these sources offers great insight to guide designing models that better handle selection bias. This paper is an attempt to conceptualize this line of thought and provide a path to explore it further.
In this work, we propose an algorithm to (1) identify disentangled representations of the above-mentioned underlying factors from any given observational datase

t D and (2) leverage this knowledge to reduce, as well as account for, the negat ive impact of selection bias on estimating the treatment effects from D. Our emp irical results show that the proposed method achieves state-of-the-art performan ce in both individual and population based evaluation measures.
****

Learning relevant features for statistical inference
Cédric Bény
We introduce an new technique to learn correlations between two types of data. The learned representation can be used to directly compute the expectations of f unctions over one type of data conditioned on the other, such as Bayesian estima tors and their standard deviations.
Specifically, our loss function teaches two neural nets to extract features repr esenting the probability vectors of highest singular value for the stochastic ma p (set of conditional probabilities) implied by the joint dataset, relative to t he inner product defined by the Fisher information metrics evaluated at the marg inals.
We test the approach using a synthetic dataset, analytical calculations, and inf erence on occluded MNIST images.
Surprisingly, when applied to supervised learning (one dataset consists of label s), this approach automatically provides regularization and faster convergence c ompared to the cross-entropy objective.
We also explore using this approach to discover salient independent features of a single dataset.
****

VILD: Variational Imitation Learning with Diverse-quality Demonstrations
Voot Tangkaratt,Bo Han,Mohammad Emtiyaz Khan,Masashi Sugiyama
The goal of imitation learning (IL) is to learn a good policy from high-quality demonstrations. However, the quality of demonstrations in reality can be diverse , since it is easier and cheaper to collect demonstrations from a mix of experts and amateurs. IL in such situations can be challenging, especially when the lev el of demonstrators' expertise is unknown. We propose a new IL paradigm called V ariational Imitation Learning with Diverse-quality demonstrations (VILD), where we explicitly model the level of demonstrators' expertise with a probabilistic g raphical model and estimate it along with a reward function. We show that a naiv e estimation approach is not suitable to large state and action spaces, and fix this issue by using a variational approach that can be easily implemented using existing reinforcement learning methods. Experiments on continuous-control bench marks demonstrate that VILD outperforms state-of-the-art methods. Our work enabl es scalable and data-efficient IL under more realistic settings than before.
****

Entropy Minimization In Emergent Languages
Eugene Kharitonov,Rahma Chaabouni,Diane Bouchacourt,Marco Baroni
There is a growing interest in studying the languages emerging when neural agent s are jointly trained to solve tasks requiring communication through a discrete channel. We investigate here the information-theoretic complexity of such langu ages, focusing on the basic two-agent, one-exchange setup. We find that, under c ommon training procedures, the emergent languages are subject to an entropy mini mization pressure that has also been detected in human language, whereby the mut ual information between the communicating agent's inputs and the messages is min imized, within the range afforded by the need for successful communication. This pressure is amplified as we increase communication channel discreteness. Furthe r, we observe that stronger discrete-channel-driven entropy minimization leads t o representations with increased robustness to overfitting and adversarial attac ks. We conclude by discussing the implications of our findings for the study of natural and artificial communication systems.
****

A Unified framework for randomized smoothing based certified defenses
Tianhang Zheng,Di Wang,Baochun Li,Jinhui Xu
Randomized smoothing, which was recently proved to be a certified defensive tech nique, has received considerable attention due to its scalability to large datas

ets and neural networks. However, several important questions still remain unans
wered in the existing frameworks, such as (i) whether Gaussian mechanism is an o
ptimal choice for certifying ℓ_2-normed robustness, and (ii) whether random
ized smoothing can certify ℓ_∞-normed robustness (on high-dimensional
datasets like ImageNet). To answer these questions, we introduce a {\em unified
} and {\em self-contained} framework to study randomized smoothing-based certifi
ed defenses, where we mainly focus on the two most popular norms in adversarial
machine learning, {\em i.e.,} ℓ_2 and ℓ_∞ norm. We answer the abo
ve two questions by first demonstrating that Gaussian mechanism and Exponential
mechanism are the (near) optimal options to certify the ℓ_2 and $\ell_\inf
ty$-normed robustness. We further show that the largest ℓ_∞ radius cer
tified by randomized smoothing is upper bounded by $O(1/\sqrt{d})$, where d is
the dimensionality of the data. This theoretical finding suggests that certifyi
ng ℓ_∞-normed robustness by randomized smoothing may not be scalable t
o high-dimensional data. The veracity of our framework and analysis is verified
by extensive evaluations on CIFAR10 and ImageNet.
****

Analysis of Video Feature Learning in Two-Stream CNNs on the Example of Zebrafis
h Swim Bout Classification
Bennet Breier,Arno Onken
Semmelhack et al. (2014) have achieved high classification accuracy in distingui
shing swim bouts of zebrafish using a Support Vector Machine (SVM). Convolutiona
l Neural Networks (CNNs) have reached superior performance in various image reco
gnition tasks over SVMs, but these powerful networks remain a black box. Reachin
g better transparency helps to build trust in their classifications and makes le
arned features interpretable to experts. Using a recently developed technique ca
lled Deep Taylor Decomposition, we generated heatmaps to highlight input regions
of high relevance for predictions. We find that our CNN makes predictions by an
alyzing the steadiness of the tail's trunk, which markedly differs from the manu
ally extracted features used by Semmelhack et al. (2014). We further uncovered t
hat the network paid attention to experimental artifacts. Removing these artifac
ts ensured the validity of predictions. After correction, our best CNN beats the
SVM by 6.12%, achieving a classification accuracy of 96.32%. Our work thus demo
nstrates the utility of AI explainability for CNNs.
****

MIST: Multiple Instance Spatial Transformer Networks
Baptiste Angles,Simon Kornblith,Shahram Izadi,Andrea Tagliasacchi,Kwang Moo Yi
We propose a deep network that can be trained to tackle image reconstruction and
classification problems that involve detection of multiple object instances, wi
thout any supervision regarding their whereabouts. The network learns to extract
the most significant top-K patches, and feeds these patches to a task-specific
network -- e.g., auto-encoder or classifier -- to solve a domain specific proble
m. The challenge in training such a network is the non-differentiable top-K sele
ction process. To address this issue, we lift the training optimization problem
by treating the result of top-K selection as a slack variable, resulting in a si
mple, yet effective, multi-stage training. Our method is able to learn to detect
recurrent structures in the training dataset by learning to reconstruct images.
It can also learn to localize structures when only knowledge on the occurrence
of the object is provided, and in doing so it outperforms the state-of-the-art.
****

ISBNet: Instance-aware Selective Branching Networks
Shaofeng Cai,Yao Shu,Wei Wang,Gang Chen,Beng Chin Ooi
Recent years have witnessed growing interests in designing efficient neural netw
orks and neural architecture search (NAS). Although remarkable efficiency and ac
curacy have been achieved, existing expert designed and NAS models neglect the f
act that input instances are of varying complexity and thus different amounts of
computation are required. Inference with a fixed model that processes all insta
nces through the same transformations would incur computational resources unnece
ssarily. Customizing the model capacity in an instance-aware manner is required
to alleviate such a problem. In this paper, we propose a novel Instance-aware Se

lective Branching Network-ISBNet to support efficient instance-level inference b
y selectively bypassing transformation branches of insignificant importance weig
ht. These weights are dynamically determined by a lightweight hypernetwork Selec
tionNet and recalibrated by gumbel-softmax for sparse branch selection. Extensiv
e experiments show that ISBNet achieves extremely efficient inference in terms o
f parameter size and FLOPs comparing to existing networks. For example, ISBNet t
akes only 8.70% parameters and 31.01% FLOPs of the efficient network MobileNetV2
 with comparable accuracy on CIFAR-10.
****

MODiR: Multi-Objective Dimensionality Reduction for Joint Data Visualisation
Tim Repke,Ralf Krestel
Many large text collections exhibit graph structures, either inherent to the con
tent itself or encoded in the metadata of the individual documents.
Example graphs extracted from document collections are co-author networks, citat
ion networks, or named-entity-cooccurrence networks.
Furthermore, social networks can be extracted from email corpora, tweets, or soc
ial media.
When it comes to visualising these large corpora, either the textual content or
the network graph are used.

In this paper, we propose to incorporate both, text and graph, to not only visua
lise the semantic information encoded in the documents' content but also the rel
ationships expressed by the inherent network structure.
To this end, we introduce a novel algorithm based on multi-objective optimisatio
n to jointly position embedded documents and graph nodes in a two-dimensional la
ndscape.
We illustrate the effectiveness of our approach with real-world datasets and sho
w that we can capture the semantics of large document collections better than ot
her visualisations based on either the content or the network information.
****

Robust Local Features for Improving the Generalization of Adversarial Training
Chuanbiao Song,Kun He,Jiadong Lin,Liwei Wang,John E. Hopcroft
Adversarial training has been demonstrated as one of the most effective methods
for training robust models to defend against adversarial examples. However, adve
rsarially trained models often lack adversarially robust generalization on unsee
n testing data. Recent works show that adversarially trained models are more bia
sed towards global structure features. Instead, in this work, we would like to i
nvestigate the relationship between the generalization of adversarial training a
nd the robust local features, as the robust local features generalize well for u
nseen shape variation. To learn the robust local features, we develop a Random B
lock Shuffle (RBS) transformation to break up the global structure features on n
ormal adversarial examples. We continue to propose a new approach called Robust
Local Features for Adversarial Training (RLFAT), which first learns the robust l
ocal features by adversarial training on the RBS-transformed adversarial example
s, and then transfers the robust local features into the training of normal adve
rsarial examples. To demonstrate the generality of our argument, we implement RL
FAT in currently state-of-the-art adversarial training frameworks. Extensive exp
eriments on STL-10, CIFAR-10 and CIFAR-100 show that RLFAT significantly improve
s both the adversarially robust generalization and the standard generalization o
f adversarial training. Additionally, we demonstrate that our models capture mor
e local features of the object on the images, aligning better with human percept
ion.
****

Online and stochastic optimization beyond Lipschitz continuity: A Riemannian app
roach
Kimon Antonakopoulos,E. Veronica Belmega,Panayotis Mertikopoulos
Motivated by applications to machine learning and imaging science, we study a cl
ass of online and stochastic optimization problems with loss functions that are
not Lipschitz continuous; in particular, the loss functions encountered by the o
ptimizer could exhibit gradient singularities or be singular themselves. Drawing

on tools and techniques from Riemannian geometry, we examine a Riemann–Lipschitz (RL) continuity condition which is tailored to the singularity landscape of the problem’s loss functions. In this way, we are able to tackle cases beyond the Lipschitz framework provided by a global norm, and we derive optimal regret bounds and last iterate convergence results through the use of regularized learning methods (such as online mirror descent). These results are subsequently validated in a class of stochastic Poisson inverse problems that arise in imaging science.

****

Distributed Online Optimization with Long-Term Constraints

Deming Yuan,Alexandre Proutiere,Guodong Shi

We consider distributed online convex optimization problems, where the distributed system consists of various computing units connected through a time-varying communication graph. In each time step, each computing unit selects a constrained vector, experiences a loss equal to an arbitrary convex function evaluated at this vector, and may communicate to its neighbors in the graph. The objective is to minimize the system-wide loss accumulated over time. We propose a decentralized algorithm with regret and cumulative constraint violation in ${\cal O}(T^{\max\{c,1-c\} })$ and ${\cal O}(T^{1-c/2})$, respectively, for any $c\in (0,1)$, where T is the time horizon. When the loss functions are strongly convex, we establish improved regret and constraint violation upper bounds in ${\cal O}(\log(T))$ and ${\cal O}(\sqrt{T\log(T)})$. These regret scalings match those obtained by state-of-the-art algorithms and fundamental limits in the corresponding centralized online optimization problem (for both convex and strongly convex loss functions). In the case of bandit feedback, the proposed algorithms achieve a regret and constraint violation in ${\cal O}(T^{\max\{c,1-c/3 \} })$ and ${\cal O}(T^{1-c/2})$ for any $c\in (0,1)$. We numerically illustrate the performance of our algorithms for the particular case of distributed online regularized linear regression problems.

****

Reinforcement Learning with Competitive Ensembles of Information-Constrained Primitives

Anirudh Goyal,Shagun Sodhani,Jonathan Binas,Xue Bin Peng,Sergey Levine,Yoshua Bengio

Reinforcement learning agents that operate in diverse and complex environments can benefit from the structured decomposition of their behavior. Often, this is addressed in the context of hierarchical reinforcement learning, where the aim is to decompose a policy into lower-level primitives or options, and a higher-level meta-policy that triggers the appropriate behaviors for a given situation. However, the meta-policy must still produce appropriate decisions in all states.

In this work, we propose a policy design that decomposes into primitives, similarly to hierarchical reinforcement learning, but without a high-level meta-policy. Instead, each primitive can decide for themselves whether they wish to act in the current state.

We use an information-theoretic mechanism for enabling this decentralized decision: each primitive chooses how much information it needs about the current state to make a decision and the primitive that requests the most information about the current state acts in the world. The primitives are regularized to use as little information as possible, which leads to natural competition and specialization. We experimentally demonstrate that this policy architecture improves over both flat and hierarchical policies in terms of generalization.

****

Learning the Arrow of Time for Problems in Reinforcement Learning

Nasim Rahaman,Steffen Wolf,Anirudh Goyal,Roman Remme,Yoshua Bengio

We humans have an innate understanding of the asymmetric progression of time, which we use to efficiently and safely perceive and manipulate our environment. Drawing inspiration from that, we approach the problem of learning an arrow of time in a Markov (Decision) Process. We illustrate how a learned arrow of time can capture salient information about the environment, which in turn can be used to measure reachability, detect side-effects and to obtain an intrinsic reward sign

al. Finally, we propose a simple yet effective algorithm to parameterize the pro
blem at hand and learn an arrow of time with a function approximator (here, a de
ep neural network). Our empirical results span a selection of discrete and conti
nuous environments, and demonstrate for a class of stochastic processes that the
 learned arrow of time agrees reasonably well with a well known notion of an arr
ow of time due to Jordan, Kinderlehrer and Otto (1998).
****

The Variational Bandwidth Bottleneck: Stochastic Evaluation on an Information Bu
dget
Anirudh Goyal,Yoshua Bengio,Matthew Botvinick,Sergey Levine
In many applications, it is desirable to extract only the relevant information f
rom complex input data, which involves making a decision about which input featu
res are relevant.
The information bottleneck method formalizes this as an information-theoretic op
timization problem by maintaining an optimal tradeoff between compression (throw
ing away irrelevant input information), and predicting the target. In many probl
em settings, including the reinforcement learning problems we consider in this w
ork, we might prefer to compress only part of the input. This is typically the c
ase when we have a standard conditioning input, such as a state observation, and
 a ``privileged'' input, which might correspond to the goal of a task, the outpu
t of a costly planning algorithm, or communication with another agent. In such c
ases, we might prefer to compress the privileged input, either to achieve better
 generalization (e.g., with respect to goals) or to minimize access to costly in
formation (e.g., in the case of communication). Practical implementations of the
 information bottleneck based on variational inference require access to the pri
vileged input in order to compute the bottleneck variable, so although they perf
orm compression, this compression operation itself needs unrestricted, lossless
access. In this work, we propose the variational bandwidth bottleneck, which dec
ides for each example on the estimated value of the privileged information befor
e seeing it, i.e., only based on the standard input, and then accordingly choose
s stochastically, whether to access the privileged input or not. We formulate a
tractable approximation to this framework and demonstrate in a series of reinfor
cement learning experiments that it can improve generalization and reduce access
 to computationally costly information.
****

AutoGrow: Automatic Layer Growing in Deep Convolutional Networks
Wei Wen,Feng Yan,Hai Li
Depth is a key component of Deep Neural Networks (DNNs), however, designing dept
h is heuristic and requires many human efforts. We propose AutoGrow to automate
depth discovery in DNNs: starting from a shallow seed architecture, AutoGrow gro
ws new layers if the growth improves the accuracy; otherwise, stops growing and
thus discovers the depth. We propose robust growing and stopping policies to gen
eralize to different network architectures and datasets. Our experiments show th
at by applying the same policy to different network architectures, AutoGrow can
always discover near-optimal depth on various datasets of MNIST, FashionMNIST, S
VHN, CIFAR10, CIFAR100 and ImageNet. For example, in terms of accuracy-computati
on trade-off, AutoGrow discovers a better depth combination in ResNets than huma
n experts. Our AutoGrow is efficient. It discovers depth within similar time of
training a single DNN.
****

Sequence-level Intrinsic Exploration Model for Partially Observable Domains
Haiyan Yin,Jianda Chen,Sinno Jialin Pan
Training reinforcement learning policies in partially observable domains with sp
arse reward signal is an important and open problem for the research community.
In this paper, we introduce a new sequence-level intrinsic novelty model to tack
le the challenge of training reinforcement learning policies in sparse rewarded
partially observable domains. First, we propose a new reasoning paradigm to infe
r the novelty for the partially observable states, which is built upon forward d
ynamics prediction. Different from conventional approaches that perform self-pre
diction or one-step forward prediction, our proposed approach engages open-loop

multi-step prediction, which enables the difficulty of novelty prediction to flexibly scale and thus results in high-quality novelty scores. Second, we propose a novel dual-LSTM architecture to facilitate the sequence-level reasoning over the partially observable state space. Our proposed architecture efficiently synthesizes information from an observation sequence and an action sequence to derive meaningful latent representations for inferring the novelty for states. To evaluate the efficiency of our proposed approach, we conduct extensive experiments on several challenging 3D navigation tasks from ViZDoom and DeepMind Lab. We also present results on two hard-exploration domains from Atari 2600 series in Appendix to demonstrate our proposed approach could generalize beyond partially observable navigation tasks. Overall, the experiment results reveal that our proposed intrinsic novelty model could outperform several state-of-the-art curiosity baselines with considerable significance in the testified domains.
****

Pipelined Training with Stale Weights of Deep Convolutional Neural Networks
Lifu Zhang,Tarek S. Abdelrahman
The growth in the complexity of Convolutional Neural Networks (CNNs) is increasing interest in partitioning a network across multiple accelerators during training and pipelining the backpropagation computations over the accelerators. Existing approaches avoid or limit the use of stale weights through techniques such as micro-batching or weight stashing. These techniques either underutilize of accelerators or increase memory footprint. We explore the impact of stale weights on the statistical efficiency and performance in a pipelined backpropagation scheme that maximizes accelerator utilization and keeps memory overhead modest. We use 4 CNNs (LeNet-5, AlexNet, VGG and ResNet) and show that when pipelining is limited to early layers in a network, training with stale weights converges and results in models with comparable inference accuracies to those resulting from non-pipelined training on MNIST and CIFAR-10 datasets; a drop in accuracy of 0.4%, 4%, 0.83% and 1.45% for the 4 networks, respectively. However, when pipelining is deeper in the network, inference accuracies drop significantly. We propose combining pipelined and non-pipelined training in a hybrid scheme to address this drop. We demonstrate the implementation and performance of our pipelined backpropagation in PyTorch on 2 GPUs using ResNet, achieving speedups of up to 1.8X over a 1-GPU baseline, with a small drop in inference accuracy.
****

Universal Learning Approach for Adversarial Defense
Uriya Pesso,Koby Bibas,Meir Feder
Adversarial attacks were shown to be very effective in degrading the performance of neural networks. By slightly modifying the input, an almost identical input is misclassified by the network. To address this problem, we adopt the universal learning framework. In particular, we follow the recently suggested Predictive Normalized Maximum Likelihood (pNML) scheme for universal learning, whose goal is to optimally compete with a reference learner that knows the true label of the test sample but is restricted to use a learner from a given hypothesis class. In our case, the reference learner is using his knowledge on the true test label to perform minor refinements to the adversarial input. This reference learner achieves perfect results on any adversarial input. The proposed strategy is designed to be as close as possible to the reference learner in the worst-case scenario. Specifically, the defense essentially refines the test data according to the different hypotheses, where each hypothesis assumes a different label for the sample. Then by comparing the resulting hypotheses probabilities, we predict the label and detect whether the sample is adversarial or natural. Combining our method with adversarial training we create a robust scheme which can handle adversarial input along with detection of the attack. The resulting scheme is demonstrated empirically.
****

The Implicit Bias of Depth: How Incremental Learning Drives Generalization
Daniel Gissin,Shai Shalev-Shwartz,Amit Daniely
A leading hypothesis for the surprising generalization of neural networks is that the dynamics of gradient descent bias the model towards simple solutions, by s

earching through the solution space in an incremental order of complexity. We fo
rmally define the notion of incremental learning dynamics and derive the conditi
ons on depth and initialization for which this phenomenon arises in deep linear
models. Our main theoretical contribution is a dynamical depth separation result
, proving that while shallow models can exhibit incremental learning dynamics, t
hey require the initialization to be exponentially small for these dynamics to p
resent themselves. However, once the model becomes deeper, the dependence become
s polynomial and incremental learning can arise in more natural settings. We com
plement our theoretical findings by experimenting with deep matrix sensing, quad
ratic neural networks and with binary classification using diagonal and convolut
ional linear networks, showing all of these models exhibit incremental learning.
****

Rethinking Softmax Cross-Entropy Loss for Adversarial Robustness
Tianyu Pang,Kun Xu,Yinpeng Dong,Chao Du,Ning Chen,Jun Zhu
Previous work shows that adversarially robust generalization requires larger sam
ple complexity, and the same dataset, e.g., CIFAR-10, which enables good standar
d accuracy may not suffice to train robust models. Since collecting new training
data could be costly, we focus on better utilizing the given data by inducing t
he regions with high sample density in the feature space, which could lead to lo
cally sufficient samples for robust learning. We first formally show that the so
ftmax cross-entropy (SCE) loss and its variants convey inappropriate supervisory
signals, which encourage the learned feature points to spread over the space sp
arsely in training. This inspires us to propose the Max-Mahalanobis center (MMC)
loss to explicitly induce dense feature regions in order to benefit robustness.
Namely, the MMC loss encourages the model to concentrate on learning ordered an
d compact representations, which gather around the preset optimal centers for di
fferent classes. We empirically demonstrate that applying the MMC loss can signi
ficantly improve robustness even under strong adaptive attacks, while keeping st
ate-of-the-art accuracy on clean inputs with little extra computation compared t
o the SCE loss.
****

Measuring Compositional Generalization: A Comprehensive Method on Realistic Data
Daniel Keysers,Nathanael Schärli,Nathan Scales,Hylke Buisman,Daniel Furrer,Sergi
i Kashubin,Nikola Momchev,Danila Sinopalnikov,Lukasz Stafiniak,Tibor Tihon,Dmitr
y Tsarkov,Xiao Wang,Marc van Zee,Olivier Bousquet
State-of-the-art machine learning methods exhibit limited compositional generali
zation. At the same time, there is a lack of realistic benchmarks that comprehen
sively measure this ability, which makes it challenging to find and evaluate imp
rovements. We introduce a novel method to systematically construct such benchmar
ks by maximizing compound divergence while guaranteeing a small atom divergence
between train and test sets, and we quantitatively compare this method to other
approaches for creating compositional generalization benchmarks. We present a la
rge and realistic natural language question answering dataset that is constructe
d according to this method, and we use it to analyze the compositional generaliz
ation ability of three machine learning architectures. We find that they fail to
generalize compositionally and that there is a surprisingly strong negative cor
relation between compound divergence and accuracy. We also demonstrate how our m
ethod can be used to create new compositionality benchmarks on top of the existi
ng SCAN dataset, which confirms these findings.

****

Theory and Evaluation Metrics for Learning Disentangled Representations
Kien Do,Truyen Tran
We make two theoretical contributions to disentanglement learning by (a) definin
g precise semantics of disentangled representations, and (b) establishing robust
metrics for evaluation. First, we characterize the concept “disentangled repres
entations” used in supervised and unsupervised methods along three dimensions–in
formativeness, separability and interpretability–which can be expressed and quan
tified explicitly using information-theoretic constructs. This helps explain the
behaviors of several well-known disentanglement learning models. We then propos

e robust metrics for measuring informativeness, separability and interpretabilit y. Through a comprehensive suite of experiments, we show that our metrics correc tly characterize the representations learned by different methods and are consis tent with qualitative (visual) results. Thus, the metrics allow disentanglement learning methods to be compared on a fair ground. We also empirically uncovered new interesting properties of VAE-based methods and interpreted them with our fo rmulation. These findings are promising and hopefully will encourage the design of more theoretically driven models for learning disentangled representations.
****

Mixup Inference: Better Exploiting Mixup to Defend Adversarial Attacks
Tianyu Pang*,Kun Xu*,Jun Zhu
It has been widely recognized that adversarial examples can be easily crafted to fool deep networks, which mainly root from the locally non-linear behavior near by input examples. Applying mixup in training provides an effective mechanism to improve generalization performance and model robustness against adversarial per turbations, which introduces the globally linear behavior in-between training ex amples. However, in previous work, the mixup-trained models only passively defen d adversarial attacks in inference by directly classifying the inputs, where the induced global linearity is not well exploited. Namely, since the locality of t he adversarial perturbations, it would be more efficient to actively break the l ocality via the globality of the model predictions. Inspired by simple geometric intuition, we develop an inference principle, named mixup inference (MI), for m ixup-trained models. MI mixups the input with other random clean samples, which can shrink and transfer the equivalent perturbation if the input is adversarial. Our experiments on CIFAR-10 and CIFAR-100 demonstrate that MI can further impro ve the adversarial robustness for the models trained by mixup and its variants.
****

Dynamically Pruned Message Passing Networks for Large-scale Knowledge Graph Reas oning
Xiaoran Xu,Wei Feng,Yunsheng Jiang,Xiaohui Xie,Zhiqing Sun,Zhi-Hong Deng
We propose Dynamically Pruned Message Passing Networks (DPMPN) for large-scale k nowledge graph reasoning. In contrast to existing models, embedding-based or pat h-based, we learn an input-dependent subgraph to explicitly model a sequential r easoning process. Each subgraph is dynamically constructed, expanding itself sel ectively under a flow-style attention mechanism. In this way, we can not only co nstruct graphical explanations to interpret prediction, but also prune message p assing in Graph Neural Networks (GNNs) to scale with the size of graphs. We take the inspiration from the consciousness prior proposed by Bengio to design a two -GNN framework to encode global input-invariant graph-structured representation and learn local input-dependent one coordinated by an attention module. Experime nts show the reasoning capability in our model that is providing a clear graphic al explanation as well as predicting results accurately, outperforming most stat e-of-the-art methods in knowledge base completion tasks.
****

A TWO-STAGE FRAMEWORK FOR MATHEMATICAL EXPRESSION RECOGNITION
Jin Zhang,Weipeng Ming,Pengfei Liu

Although mathematical expressions (MEs) recognition have achieved great progress , the development of MEs recognition in real scenes is still unsatisfactory. Ins pired by the recent work of neutral network, this paper proposes a novel two-sta ge approach which takes a printed mathematical expression image as input and gen erates LaTeX sequence as output. In the first stage, this method locates and rec ognizes the math symbols of input image by object detection algorithm. In the se cond stage, it translates math symbols with position information into LaTeX sequ ences by seq2seq model equipped with attention mechanism. In particular, the det ection of mathematical symbols and the structural analysis of mathematical formu las are carried out separately in two steps, which effectively improves the reco gnition accuracy and enhances the generalization ability. The experiment demonst rates that the two-stage method significantly outperforms the end-to-end method. Especially, the ExpRate(expression recognition rate) of our model is 74.1%, 20.

3 percentage points higher than that of the end-to-end model on the test data th
at doesn’t come from the same source as training data.
****
Learning Invariants through Soft Unification
Nuri Cingillioglu,Alessandra Russo
Human reasoning involves recognising common underlying principles across many ex
amples by utilising variables. The by-products of such reasoning are invariants
that capture patterns across examples such as "if someone went somewhere then th
ey are there" without mentioning specific people or places. Humans learn what va
riables are and how to use them at a young age, and the question this paper addr
esses is whether machines can also learn and use variables solely from examples
without requiring human pre-engineering. We propose Unification Networks that in
corporate soft unification into neural networks to learn variables and by doing
so lift examples into invariants that can then be used to solve a given task. We
 evaluate our approach on four datasets to demonstrate that learning invariants
captures patterns in the data and can improve performance over baselines.
****
Are Pre-trained Language Models Aware of Phrases? Simple but Strong Baselines fo
r Grammar Induction
Taeuk Kim,Jihun Choi,Daniel Edmiston,Sang-goo Lee
With the recent success and popularity of pre-trained language models (LMs) in n
atural language processing, there has been a rise in efforts to understand their
 inner workings.
In line with such interest, we propose a novel method that assists us in investi
gating the extent to which pre-trained LMs capture the syntactic notion of const
ituency.
Our method provides an effective way of extracting constituency trees from the p
re-trained LMs without training.
In addition, we report intriguing findings in the induced trees, including the f
act that pre-trained LMs outperform other approaches in correctly demarcating ad
verb phrases in sentences.
****
FSPool: Learning Set Representations with Featurewise Sort Pooling
Yan Zhang,Jonathon Hare,Adam Prügel-Bennett
Traditional set prediction models can struggle with simple datasets due to an is
sue we call the responsibility problem. We introduce a pooling method for sets o
f feature vectors based on sorting features across elements of the set. This can
 be used to construct a permutation-equivariant auto-encoder that avoids this re
sponsibility problem. On a toy dataset of polygons and a set version of MNIST, w
e show that such an auto-encoder produces considerably better reconstructions an
d representations. Replacing the pooling function in existing set encoders with
FSPool improves accuracy and convergence speed on a variety of datasets.
****
Recurrent Neural Networks are Universal Filters
Wenjie Xu,Xiuqiong Chen,Stephen S.-T. Yau
Recurrent neural networks (RNN) are powerful time series modeling tools in ma-
chine learning. It has been successfully applied in a variety of fields such as
natural
language processing (Mikolov et al. (2010), Graves et al. (2013), Du et al. (201
5)),
control (Fei & Lu (2017)) and traffic forecasting (Ma et al. (2015)), etc. In th
ose
application scenarios, RNN can be viewed as implicitly modelling a stochastic dy
-
namic system. Another type of popular neural network, deep (feed-forward) neural
network has also been successfully applied in different engineering disciplines,
whose approximation capability has been well characterized by universal approxi-
mation theorem (Hornik et al. (1989), Park & Sandberg (1991), Lu et al. (2017)).
However, the underlying approximation capability of RNN has not been fully
understood in a quantitative way. In our paper, we consider a stochastic dynamic

system with noisy observations and analyze the approximation capability of RNN in synthesizing the optimal state estimator, namely optimal filter. We unify the recurrent neural network into Bayesian filtering framework and show that recurrent neural network is a universal approximator of optimal finite dimensional filters under some mild conditions. That is to say, for any stochastic dynamic systems with noisy sequential observations that satisfy some mild conditions, we show that (informal) ∀■ > 0, ∃ RNN-based filter, s.t. lim sup x■ k|k − E[x k |Y k] < ■, k→∞ where x■ k|k is RNN-based filter’s estimate of state x k at step k conditioned on the observation history and E[x k |Y k] is the conditional mean of x k , known as the optimal estimate of the state in minimum mean square error sense. As an interesting special case, the widely used Kalman filter (KF) can be synthesized by RNN.

****

On the Convergence of FedAvg on Non-IID Data

Xiang Li,Kaixuan Huang,Wenhao Yang,Shusen Wang,Zhihua Zhang

Federated learning enables a large amount of edge computing devices to jointly learn a model without data sharing. As a leading algorithm in this setting, Federated Averaging (\texttt{FedAvg}) runs Stochastic Gradient Descent (SGD) in parallel on a small subset of the total devices and averages the sequences only once in a while. Despite its simplicity, it lacks theoretical guarantees under realistic settings. In this paper, we analyze the convergence of \texttt{FedAvg} on non-iid data and establish a convergence rate of $\mathcal{O}(\frac{1}{T})$ for strongly convex and smooth problems, where T is the number of SGDs. Importantly, our bound demonstrates a trade-off between communication-efficiency and convergence rate. As user devices may be disconnected from the server, we relax the assumption of full device participation to partial device participation and study different averaging schemes; low device participation rate can be achieved without severely slowing down the learning. Our results indicate that heterogeneity of data slows down the convergence, which matches empirical observations. Furthermore, we provide a necessary condition for \texttt{FedAvg} on non-iid data: the learning rate η must decay, even if full-gradient is used; otherwise, the solution will be $\Omega (\eta)$ away from the optimal.

****

Adversarially Robust Neural Networks via Optimal Control: Bridging Robustness with Lyapunov Stability

Zhiyang Chen,Hang Su

Deep neural networks are known to be vulnerable to adversarial perturbations. In this paper, we bridge adversarial robustness of neural nets with Lyapunov stability of dynamical systems. From this viewpoint, training neural nets is equivalent to finding an optimal control of the discrete dynamical system, which allows one to utilize methods of successive approximations, an optimal control algorithm based on Pontryagin's maximum principle, to train neural nets. This decoupled training method allows us to add constraints to the optimization, which makes the deep model more robust. The constrained optimization problem can be formulated as a semi-definite programming problem and hence can be solved efficiently. Experiments show that our method effectively improves deep model's adversarial robustness.

****

Multi-agent Reinforcement Learning for Networked System Control

Tianshu Chu,Sandeep Chinchali,Sachin Katti

This paper considers multi-agent reinforcement learning (MARL) in networked system control. Specifically, each agent learns a decentralized control policy based on local observations and messages from connected neighbors. We formulate such a networked MARL (NMARL) problem as a spatiotemporal Markov decision process and

introduce a spatial discount factor to stabilize the training of each local agent. Further, we propose a new differentiable communication protocol, called NeurComm, to reduce information loss and non-stationarity in NMARL. Based on experiments in realistic NMARL scenarios of adaptive traffic signal control and cooperative adaptive cruise control, an appropriate spatial discount factor effectively enhances the learning curves of non-communicative MARL algorithms, while NeurComm outperforms existing communication protocols in both learning efficiency and control performance.

****

Learning to Anneal and Prune Proximity Graphs for Similarity Search

Minjia Zhang,Wenhan Wang,Yuxiong He

This paper studies similarity search, which is a crucial enabler of many feature vector--based applications. The problem of similarity search has been extensively studied in the machine learning community. Recent advances of proximity graphs have achieved outstanding performance through exploiting the navigability of the underlying graph structure. In this work, we introduce the annealable proximity graph (APG) method to learn and reshape proximity graphs for efficiency and effective similarity search. APG makes proximity graph edges annealable, which can be effectively trained with a stochastic optimization algorithm. APG identifies important edges that best preserve graph navigability and prune inferior edges without drastically changing graph properties. Experimental results show that APG achieves state-of-the-art results not only by producing proximity graphs with less number of edges but also speeding up the search time by 20--40\% across different datasets with almost no loss of accuracy.

****

Deep Bayesian Structure Networks

Zhijie Deng,Yucen Luo,Jun Zhu,Bo Zhang

Bayesian neural networks (BNNs) introduce uncertainty estimation to deep networks by performing Bayesian inference on network weights. However, such models bring the challenges of inference, and further BNNs with weight uncertainty rarely achieve superior performance to standard models. In this paper, we investigate a new line of Bayesian deep learning by performing Bayesian reasoning on the structure of deep neural networks. Drawing inspiration from the neural architecture search, we define the network structure as random weights on the redundant operations between computational nodes, and apply stochastic variational inference techniques to learn the structure distributions of networks. Empirically, the proposed method substantially surpasses the advanced deep neural networks across a range of classification and segmentation tasks. More importantly, our approach also preserves benefits of Bayesian principles, producing improved uncertainty estimation than the strong baselines including MC dropout and variational BNNs algorithms (e.g. noisy EK-FAC).

****

Hierarchical Foresight: Self-Supervised Learning of Long-Horizon Tasks via Visual Subgoal Generation

Suraj Nair,Chelsea Finn

Video prediction models combined with planning algorithms have shown promise in enabling robots to learn to perform many vision-based tasks through only self-supervision, reaching novel goals in cluttered scenes with unseen objects. However, due to the compounding uncertainty in long horizon video prediction and poor scalability of sampling-based planning optimizers, one significant limitation of these approaches is the ability to plan over long horizons to reach distant goals. To that end, we propose a framework for subgoal generation and planning, hierarchical visual foresight (HVF), which generates subgoal images conditioned on a goal image, and uses them for planning. The subgoal images are directly optimized to decompose the task into easy to plan segments, and as a result, we observe that the method naturally identifies semantically meaningful states as subgoals. Across three out of four simulated vision-based manipulation tasks, we find that our method achieves more than 20% absolute performance improvement over planning without subgoals and model-free RL approaches. Further, our experiments illu

strate that our approach extends to real, cluttered visual scenes.
****
Keyframing the Future: Discovering Temporal Hierarchy with Keyframe-Inpainter Prediction
Karl Pertsch,Oleh Rybkin,Jingyun Yang,Konstantinos G. Derpanis,Kostas Daniilidis,Joseph J. Lim,Andrew Jaegle
To flexibly and efficiently reason about temporal sequences, abstract representations that compactly represent the important information in the sequence are needed. One way of constructing such representations is by focusing on the important events in a sequence. In this paper, we propose a model that learns both to discover such key events (or keyframes) as well as to represent the sequence in terms of them. We do so using a hierarchical Keyframe-Inpainter (KeyIn) model that first generates keyframes and their temporal placement and then inpaints the sequences between keyframes. We propose a fully differentiable formulation for efficiently learning the keyframe placement. We show that KeyIn finds informative keyframes in several datasets with diverse dynamics. When evaluated on a planning task, KeyIn outperforms other recent proposals for learning hierarchical representations.
****
Differential Privacy in Adversarial Learning with Provable Robustness
NhatHai Phan,My T. Thai,Ruoming Jin,Han Hu,Dejing Dou
In this paper, we aim to develop a novel mechanism to preserve differential privacy (DP) in adversarial learning for deep neural networks, with provable robustness to adversarial examples. We leverage the sequential composition theory in DP, to establish a new connection between DP preservation and provable robustness. To address the trade-off among model utility, privacy loss, and robustness, we design an original, differentially private, adversarial objective function, based on the post-processing property in DP, to tighten the sensitivity of our model. An end-to-end theoretical analysis and thorough evaluations show that our mechanism notably improves the robustness of DP deep neural networks.
****
Topology-Aware Pooling via Graph Attention
Hongyang Gao,Shuiwang Ji
Pooling operations have shown to be effective on various tasks in computer vision and natural language processing. One challenge of performing pooling operations on graph data is the lack of locality that is not well-defined on graphs. Previous studies used global ranking methods to sample some of the important nodes, but most of them are not able to incorporate graph topology information in computing ranking scores. In this work, we propose the topology-aware pooling (TAP) layer that uses attention operators to generate ranking scores for each node by attending each node to its neighboring nodes. The ranking scores are generated locally while the selection is performed globally, which enables the pooling operation to consider topology information. To encourage better graph connectivity in the sampled graph, we propose to add a graph connectivity term to the computation of ranking scores in the TAP layer. Based on our TAP layer, we develop a network on graph data, known as the topology-aware pooling network. Experimental results on graph classification tasks demonstrate that our methods achieve consistently better performance than previous models.
****
Siamese Attention Networks
Hongyang Gao,Yaochen Xie,Shuiwang Ji
Attention operators have been widely applied on data of various orders and dimensions such as texts, images, and videos. One challenge of applying attention operators is the excessive usage of computational resources. This is due to the usage of dot product and softmax operator when computing similarity scores. In this work, we propose the Siamese similarity function that uses a feed-forward network to compute similarity scores. This results in the Siamese attention operator (SAO). In particular, SAO leads to a dramatic reduction in the requirement of computational resources. Experimental results show that our SAO can save 94% memory usage and speed up the computation by a factor of 58 compared to the regular a

ttention operator. The computational advantage of SAO is even larger on higher-o rder and higher-dimensional data. Results on image classification and restoratio n tasks demonstrate that networks with SAOs are as effective as models with regu lar attention operator, while significantly outperform those without attention o perators.
****

Neural Stored-program Memory
Hung Le,Truyen Tran,Svetha Venkatesh
Neural networks powered with external memory simulate computer behaviors. These models, which use the memory to store data for a neural controller, can learn al gorithms and other complex tasks. In this paper, we introduce a new memory to st ore weights for the controller, analogous to the stored-program memory in modern computer architectures. The proposed model, dubbed Neural Stored-program Memory , augments current memory-augmented neural networks, creating differentiable mac hines that can switch programs through time, adapt to variable contexts and thus fully resemble the Universal Turing Machine. A wide range of experiments demons trate that the resulting machines not only excel in classical algorithmic proble ms, but also have potential for compositional, continual, few-shot learning and question-answering tasks.
****

ES-MAML: Simple Hessian-Free Meta Learning
Xingyou Song,Wenbo Gao,Yuxiang Yang,Krzysztof Choromanski,Aldo Pacchiano,Yunhao Tang
We introduce ES-MAML, a new framework for solving the model agnostic meta learni ng (MAML) problem based on Evolution Strategies (ES). Existing algorithms for MA ML are based on policy gradients, and incur significant difficulties when attemp ting to estimate second derivatives using backpropagation on stochastic policies . We show how ES can be applied to MAML to obtain an algorithm which avoids the problem of estimating second derivatives, and is also conceptually simple and ea sy to implement. Moreover, ES-MAML can handle new types of nonsmooth adaptation operators, and other techniques for improving performance and estimation of ES m ethods become applicable. We show empirically that ES-MAML is competitive with e xisting methods and often yields better adaptation with fewer queries.
****

Enforcing Physical Constraints in Neural Neural Networks through Differentiable PDE Layer
Chiyu "Max" Jiang,Karthik Kashinath,Prabhat,Philip Marcus
Recent studies at the intersection of physics and deep learning have illustrated successes in the application of deep neural networks to partially or fully repl ace costly physics simulations. Enforcing physical constraints to solutions gene rated
by neural networks remains a challenge, yet it is essential to the accuracy and trustworthiness of such model predictions. Many systems in the physical sciences are governed by Partial Differential Equations (PDEs). Enforcing these as hard constraints, we show, are inefficient in conventional frameworks due to the high dimensionality of the generated fields. To this end, we propose the use of a no vel differentiable spectral projection layer for neural networks that efficientl y enforces
spatial PDE constraints using spectral methods, yet is fully differentiable, all owing for its use as a layer in neural networks that supports end-to-end trainin g. We show that its computational cost is cheaper than a regular convolution lay er. We apply it to
an important class of physical systems – incompressible turbulent flows, where t he divergence-free PDE constraint is required. We train a 3D Conditional Generat ive Adversarial Network (CGAN) for turbulent flow super-resolution efficiently, whilst
guaranteeing the spatial PDE constraint of zero divergence. Furthermore, our emp irical results show that the model produces realistic flow fields with more accu rate flow statistics when trained with hard constraints imposed via the proposed novel differentiable spectral projection layer, as compared to soft constrained

and unconstrained counterparts.
****
TabFact: A Large-scale Dataset for Table-based Fact Verification
Wenhu Chen,Hongmin Wang,Jianshu Chen,Yunkai Zhang,Hong Wang,Shiyang Li,Xiyou Zhou,William Yang Wang
The problem of verifying whether a textual hypothesis holds based on the given evidence, also known as fact verification, plays an important role in the study of natural language understanding and semantic representation. However, existing studies are mainly restricted to dealing with unstructured evidence (e.g., natural language sentences and documents, news, etc), while verification under structured evidence, such as tables, graphs, and databases, remains unexplored. This paper specifically aims to study the fact verification given semi-structured data as evidence. To this end, we construct a large-scale dataset called TabFact with 16k Wikipedia tables as the evidence for 118k human-annotated natural language statements, which are labeled as either ENTAILED or REFUTED. TabFact is challenging since it involves both soft linguistic reasoning and hard symbolic reasoning. To address these reasoning challenges, we design two different models: Table-BERT and Latent Program Algorithm (LPA). Table-BERT leverages the state-of-the-art pre-trained language model to encode the linearized tables and statements into continuous vectors for verification. LPA parses statements into LISP-like programs and executes them against the tables to obtain the returned binary value for verification. Both methods achieve similar accuracy but still lag far behind human performance. We also perform a comprehensive analysis to demonstrate great future opportunities.
****
Evidence-Aware Entropy Decomposition For Active Deep Learning
Weishi Shi,Xujiang Zhao,Feng Chen,Qi Yu
We present a novel multi-source uncertainty prediction approach that enables deep learning (DL) models to be actively trained with much less labeled data. By leveraging the second-order uncertainty representation provided by subjective logic (SL), we conduct evidence-based theoretical analysis and formally decompose the predicted entropy over multiple classes into two distinct sources of uncertainty: vacuity and dissonance, caused by lack of evidence and conflict of strong evidence, respectively. The evidence based entropy decomposition provides deeper insights on the nature of uncertainty, which can help effectively explore a large and high-dimensional unlabeled data space. We develop a novel loss function that augments DL based evidence prediction with uncertainty anchor sample identification through kernel density estimation (KDE). The accurately estimated multiple sources of uncertainty are systematically integrated and dynamically balanced using a data sampling function for label-efficient active deep learning (ADL). Experiments conducted over both synthetic and real data and comparison with competitive AL methods demonstrate the effectiveness of the proposed ADL model.
****
Learning to Generate Grounded Visual Captions without Localization Supervision
Chih-Yao Ma,Yannis Kalantidis,Ghassan AlRegib,Peter Vajda,Marcus Rohrbach,Zsolt Kira
When automatically generating a sentence description for an image or video, it often remains unclear how well the generated caption is grounded, or if the model hallucinates based on priors in the dataset and/or the language model. The most common way of relating image regions with words in caption models is through an attention mechanism over the regions that are used as input to predict the next word. The model must therefore learn to predict the attentional weights without knowing the word it should localize. This is difficult to train without grounding supervision since recurrent models can propagate past information and there is no explicit signal to force the captioning model to properly ground the individual decoded words. In this work, we help the model to achieve this via a novel cyclical training regimen that forces the model to localize each word in the image after the sentence decoder generates it, and then reconstruct the sentence from the localized image region(s) to match the ground-truth. Our proposed framework only requires learning one extra fully-connected layer (the localizer), a lay

er that can be removed at test time. We show that our model significantly improv es grounding accuracy without relying on grounding supervision or introducing ex tra computation during inference for both image and video captioning tasks.
****
Extreme Triplet Learning: Effectively Optimizing Easy Positives and Hard Negativ es
Hong Xuan,Robert Pless
The Triplet Loss approach to Distance Metric Learning is defined by the strategy to select triplets and the loss function through which those triplets are optim ized. During optimization, two especially important cases are easy positive and hard negative mining which consider, the closest example of the same and differ ent classes. We characterize how triplets behave based during optimization as a function of these similarities, and highlight that these important cases have t echnical problems where standard gradient descent behaves poorly, pulling the ne gative example closer and/or pushing the positive example farther away. We deri ve an updated loss function that fixes these problems and shows improvements to the state of the art for CUB, CAR, SOP, In-Shop Clothes datasets.
****
Implicit Bias of Gradient Descent based Adversarial Training on Separable Data
Yan Li,Ethan X.Fang,Huan Xu,Tuo Zhao
Adversarial training is a principled approach for training robust neural network s. Despite of tremendous successes in practice, its theoretical properties still remain largely unexplored. In this paper, we provide new theoretical insights o f gradient descent based adversarial training by studying its computational prop erties, specifically on its implicit bias. We take the binary classification tas k on linearly separable data as an illustrative example, where the loss asymptot ically attains its infimum as the parameter diverges to infinity along certain d irections. Specifically, we show that for any fixed iteration T, when the adve rsarial perturbation during training has proper bounded L2 norm, the classifier learned by gradient descent based adversarial training converges in direction t o the maximum L2 norm margin classifier at the rate of $O(1/\sqrt{T})$, signific antly faster than the rate $O(1/\log T}$ of training with clean data. In additio n, when the adversarial perturbation during training has bounded Lq norm, the re sulting classifier converges in direction to a maximum mixed-norm margin classif ier, which has a natural interpretation of robustness, as being the maximum L2 n orm margin classifier under worst-case bounded Lq norm perturbation to the data. Our findings provide theoretical backups for adversarial training that it inde ed promotes robustness against adversarial perturbation.
****
Graph Warp Module: an Auxiliary Module for Boosting the Power of Graph Neural Ne tworks in Molecular Graph Analysis
Katsuhiko Ishiguro,Shin-ichi Maeda,Masanori Koyama
Graph Neural Network (GNN) is a popular architecture for the analysis of chemica l molecules, and it has numerous applications in material and medicinal science. Current lines of GNNs developed for molecular analysis, however, do not fit well on the training set, and their performance does not scale well with the complex ity of the network.
In this paper, we propose an auxiliary module to be attached to a GNN that can b oost the representation power of the model without hindering the original GNN ar chitecture.
Our auxiliary module can improve the representation power and the generalization ability of a wide variety of GNNs, including those that are used commonly in bi ochemical applications.
****
The Visual Task Adaptation Benchmark
Xiaohua Zhai,Joan Puigcerver,Alexander Kolesnikov,Pierre Ruyssen,Carlos Riquelme ,Mario Lucic,Josip Djolonga,Andre Susano Pinto,Maxim Neumann,Alexey Dosovitskiy, Lucas Beyer,Olivier Bachem,Michael Tschannen,Marcin Michalski,Olivier Bousquet,S ylvain Gelly,Neil Houlsby
Representation learning promises to unlock deep learning for the long tail of vi

sion tasks without expansive labelled datasets. Yet, the absence of a unified ya rdstick to evaluate general visual representations hinders progress. Many sub-fi elds promise representations, but each has different evaluation protocols that a re either too constrained (linear classification), limited in scope (ImageNet, C IFAR, Pascal-VOC), or only loosely related to representation quality (generation). We present the Visual Task Adaptation Benchmark (VTAB): a diverse, realistic, and challenging benchmark to evaluate representations. VTAB embodies one princi ple: good representations adapt to unseen tasks with few examples. We run a larg e VTAB study of popular algorithms, answering questions like: How effective are ImageNet representation on non-standard datasets? Are generative models competit ive? Is self-supervision useful if one already has labels?
****

Learning Similarity Metrics for Numerical Simulations
Georg Kohl,Kiwon Um,Nils Thuerey
We propose a novel approach to compute a stable and generalizing metric (LNSM) w ith convolutional neural networks (CNN) to compare field data from a variety of numerical simulation sources. Our method employs a Siamese network architecture that is motivated by the mathematical properties of a metric and is known to wor k well for finding similarities of other data modalities. We leverage a controll able data generation setup with partial differential equation (PDE) solvers to c reate increasingly different outputs from a reference simulation. In addition, t he data generation allows for adjusting the difficulty of the resulting learning task. A central component of our learned metric is a specialized loss function, that introduces knowledge about the correlation between single data samples int o the training process. To demonstrate that the proposed approach outperforms ex isting simple metrics for vector spaces and other learned, image based metrics w e evaluate the different methods on a large range of test data. Additionally, we analyze generalization benefits of using the proposed correlation loss and the impact of an adjustable training data difficulty.
****

Image-guided Neural Object Rendering
Justus Thies,Michael Zollhöfer,Christian Theobalt,Marc Stamminger,Matthias Nießn er
We propose a learned image-guided rendering technique that combines the benefits of image-based rendering and GAN-based image synthesis. The goal of our method is to generate photo-realistic re-renderings of reconstructed objects for virtua l and augmented reality applications (e.g., virtual showrooms, virtual tours and sightseeing, the digital inspection of historical artifacts). A core component of our work is the handling of view-dependent effects. Specifically, we directly train an object-specific deep neural network to synthesize the view-dependent a ppearance of an object.
As input data we are using an RGB video of the object. This video is used to rec onstruct a proxy geometry of the object via multi-view stereo. Based on this 3D proxy, the appearance of a captured view can be warped into a new target view as in classical image-based rendering. This warping assumes diffuse surfaces, in c ase of view-dependent effects, such as specular highlights, it leads to artifact s. To this end, we propose EffectsNet, a deep neural network that predicts view-dependent effects. Based on these estimations, we are able to convert observed i mages to diffuse images. These diffuse images can be projected into other views. In the target view, our pipeline reinserts the new view-dependent effects. To c omposite multiple reprojected images to a final output, we learn a composition n etwork that outputs photo-realistic results. Using this image-guided approach, t he network does not have to allocate capacity on ``remembering'' object appearan ce, instead it learns how to combine the appearance of captured images. We demon strate the effectiveness of our approach both qualitatively and quantitatively o n synthetic as well as on real data.
****

MULTIPOLAR: Multi-Source Policy Aggregation for Transfer Reinforcement Learning between Diverse Environmental Dynamics
Mohammadamin Barekatain,Ryo Yonetani,Masashi Hamaya

Transfer reinforcement learning (RL) aims at improving learning efficiency of an agent by exploiting knowledge from other source agents trained on relevant tasks. However, it remains challenging to transfer knowledge between different environmental dynamics without having access to the source environments. In this work, we explore a new challenge in transfer RL, where only a set of source policies collected under unknown diverse dynamics is available for learning a target task efficiently. To address this problem, the proposed approach, MULTI-source POLicy AggRegation (MULTIPOLAR), comprises two key techniques. We learn to aggregate the actions provided by the source policies adaptively to maximize the target task performance. Meanwhile, we learn an auxiliary network that predicts residuals around the aggregated actions, which ensures the target policy's expressiveness even when some of the source policies perform poorly. We demonstrated the effectiveness of MULTIPOLAR through an extensive experimental evaluation across six simulated environments ranging from classic control problems to challenging robotics simulations, under both continuous and discrete action spaces.
****
Effective and Robust Detection of Adversarial Examples via Benford-Fourier Coefficients
Chengcheng Ma,Baoyuan Wu,Shibiao Xu,Yanbo Fan,Yong Zhang,Xiaopeng Zhang,Zhifeng Li
Adversarial examples have been well known as a serious threat to deep neural networks (DNNs). To ensure successful and safe operations of DNNs on realworld tasks, it is urgent to equip DNNs with effective defense strategies. In this work, we study the detection of adversarial examples, based on the assumption that the output and internal responses of one DNN model for both adversarial and benign examples follow the generalized Gaussian distribution (GGD), but with different parameters (i.e., shape factor, mean, and variance). GGD is a general distribution family to cover many popular distributions (e.g., Laplacian, Gaussian, or uniform). It is more likely to approximate the intrinsic distributions of internal responses than any specific distribution. Besides, since the shape factor is more robust to different databases rather than the other two parameters, we propose to construct discriminative features via the shape factor for adversarial detection, employing the magnitude of Benford-Fourier coefficients (MBF), which can be easily estimated using responses. Finally, a support vector machine is trained as the adversarial detector through leveraging the MBF features. Through the Kolmogorov-Smirnov (KS) test, we empirically verify that: 1) the posterior vectors of both adversarial and benign examples follow GGD; 2) the extracted MBF features of adversarial and benign examples follow different distributions. Extensive experiments in terms of image classification demonstrate that the proposed detector is much more effective and robust on detecting adversarial examples of different crafting methods and different sources, in contrast to state-of-the-art adversarial detection methods.
****
Stablizing Adversarial Invariance Induction by Discriminator Matching
Yusuke Iwasawa,Kei Akuzawa,Yutaka Matsuo
Incorporating the desired invariance into representation learning is a key challenge in many situations, e.g., for domain generalization and privacy/fairness constraints. An adversarial invariance induction (AII) shows its power on this purpose, which maximizes the proxy of the conditional entropy between representations and attributes by adversarial training between an attribute discriminator and feature extractor. However, the practical behavior of AII is still unclear as the previous analysis assumes the optimality of the attribute classifier, which i

s rarely held in practice. This paper first analyzes the practical behavior of A
II both theoretically and empirically, indicating that AII has theoretical diffi
culty as it maximizes variational {\em upper} bound of the actual conditional en
tropy, and AII catastrophically fails to induce invariance even in simple cases
as suggested by the above theoretical findings. We then argue that a simple modi
fication to AII can significantly stabilize the adversarial induction framework
and achieve better invariant representations. Our modification is based on the p
roperty of conditional entropy; it is maximized if and only if the divergence be
tween all pairs of marginal distributions over z between different attributes
is minimized. The proposed method, {\em invariance induction by discriminator ma
tching}, modify AII objective to explicitly consider the divergence minimization
 requirements by defining a proxy of the divergence by using the attribute discr
iminator. Empirical validations on both the toy dataset and four real-world data
sets (related to applications of user anonymization and domain generalization) r
eveal that the proposed method provides superior performance when inducing invar
iance for nuisance factors.
****

POP-Norm: A Theoretically Justified and More Accelerated Normalization Approach
Hanyang Peng,Shiqi Yu
Batch Normalization (BatchNorm) has been a default module in modern deep networ
ks due to its effectiveness for accelerating training deep neural networks. It
 is widely accepted that the great success of BatchNorm is owing to reduction of
 internal covariate shift (ICS), but recently it is demonstrated that the link b
etween them is fairly weak. The intrinsic reason behind effectiveness of BatchNo
rm is still unrevealed that limits it to be made better use. In light of this,
we propose a new normalization approach, referred to as Pre-Operation Normaliza
tion (POP-Norm), which is theoretically ensured to speed up the training converg
ence. Not surprisingly, POP-Norm and BatchNorm are largely the same. Hence the s
imilarities can help us to theoretically interpret the root of BatchNorm's effe
ctiveness. There are still some significant distinctions between the two approac
hes. Just the distinctions make POP-Norm achieve faster convergence rate and bet
ter performance than BatchNorm, which are validated in extensive experiments on
benchmark datasets: CIFAR10, CIFAR100 and ILSVRC2012.
****

Programmable Neural Network Trojan for Pre-trained Feature Extractor
Yu Ji,Zinxin Liu,Xing Hu,Peiqi Wang,Youhui Zhang
Neural network (NN) trojaning attack is an emerging and important attack that ca
n broadly damage the system deployed with NN models.
Different from adversarial attack, it hides malicious functionality in the weigh
t parameters of NN models.
Existing studies have explored NN trojaning attacks in some small datasets for s
pecific domains, with limited numbers of fixed target classes.
In this paper, we propose a more powerful trojaning attack method for large mode
ls, which outperforms existing studies in capability, generality, and stealthine
ss.
First, the attack is programmable that the malicious misclassification target is
 not fixed and can be generated on demand even after the victim's deployment.
Second, our trojaning attack is not limited in a small domain; one trojaned mode
l on a large-scale dataset can affect applications of different domains that reu
ses its general features.
Third, our trojan shows no biased behavior for different target classes, which m
akes it more difficult to defend.
****

On Layer Normalization in the Transformer Architecture
Ruibin Xiong,Yunchang Yang,Di He,Kai Zheng,Shuxin Zheng,Huishuai Zhang,Yanyan La
n,Liwei Wang,Tie-Yan Liu
The Transformer architecture is popularly used in natural language processing ta
sks. To train a Transformer model, a carefully designed learning rate warm-up st
age is usually needed: the learning rate has to be set to an extremely small val
ue at the beginning of the optimization and then gradually increases in some giv

en number of iterations. Such a stage is shown to be crucial to the final perfor
mance and brings more hyper-parameter tunings. In this paper, we study why the l
earning rate warm-up stage is important in training the Transformer and theoreti
cally show that the location of layer normalization matters. It can be proved th
at at the beginning of the optimization, for the original Transformer, which pla
ces the layer normalization between the residual blocks, the expected gradients
of the parameters near the output layer are large. Then using a large learning r
ate on those gradients makes the training unstable. The warm-up stage is practic
ally helpful to avoid this problem. Such an analysis motivates us to investigate
a slightly modified Transformer architecture which locates the layer normalizat
ion inside the residual blocks. We show that the gradients in this Transformer a
rchitecture are well-behaved at initialization. Given these findings, we are th
e first to show that this Transformer variant is easier and faster to train. The
learning rate warm-up stage can be safely removed, and the training time can be
largely reduced on a wide range of applications.
****

PC-DARTS: Partial Channel Connections for Memory-Efficient Architecture Search
Yuhui Xu,Lingxi Xie,Xiaopeng Zhang,Xin Chen,Guo-Jun Qi,Qi Tian,Hongkai Xiong
Differentiable architecture search (DARTS) provided a fast solution in finding e
ffective network architectures, but suffered from large memory and computing ove
rheads in jointly training a super-net and searching for an optimal architecture
. In this paper, we present a novel approach, namely Partially-Connected DARTS,
by sampling a small part of super-net to reduce the redundancy in exploring the
network space, thereby performing a more efficient search without comprising th
e performance. In particular, we perform operation search in a subset of channel
s while bypassing the held out part in a shortcut. This strategy may suffer from
an undesired inconsistency on selecting the edges of super-net caused by sampli
ng different channels. We solve it by introducing edge normalization, which add
s a new set of edge-level hyper-parameters to reduce uncertainty in search. Than
ks to the reduced memory cost, PC-DARTS can be trained with a larger batch size
and, consequently, enjoy both faster speed and higher training stability. Experi
ment results demonstrate the effectiveness of the proposed method. Specifically,
we achieve an error rate of 2.57% on CIFAR10 within merely 0.1 GPU-days for arc
hitecture search, and a state-of-the-art top-1 error rate of 24.2% on ImageNet (
under the mobile setting) within 3.8 GPU-days for search. Our code has been made
available at https://www.dropbox.com/sh/on9lg3rpx1r6dkf/AABG5mt0sMHjnEJyoRnLEYW
4a?dl=0.
****

Knowledge Consistency between Neural Networks and Beyond
Ruofan Liang,Tianlin Li,Longfei Li,Jing Wang,Quanshi Zhang
This paper aims to analyze knowledge consistency between pre-trained deep neural
networks. We propose a generic definition for knowledge consistency between neu
ral networks at different fuzziness levels. A task-agnostic method is designed t
o disentangle feature components, which represent the consistent knowledge, from
raw intermediate-layer features of each neural network. As a generic tool, our
method can be broadly used for different applications. In preliminary experiment
s, we have used knowledge consistency as a tool to diagnose representations of n
eural networks. Knowledge consistency provides new insights to explain the succe
ss of existing deep-learning techniques, such as knowledge distillation and netw
ork compression. More crucially, knowledge consistency can also be used to refin
e pre-trained networks and boost performance.
****

Temporal Probabilistic Asymmetric Multi-task Learning
Nguyen Anh Tuan,Hyewon Jeong,Eunho Yang,Sungju Hwang
When performing multi-task predictions with time-series data, knowledge learned
for one task at a specific time step may be useful in learning for another task
at a later time step (e.g. prediction of sepsis may be useful for prediction of
mortality for risk prediction at intensive care units). To capture such dynamica
lly changing asymmetric relationships between tasks and long-range temporal depe
ndencies in time-series data, we propose a novel temporal asymmetric multi-task

learning model, which learns to combine features from other tasks at diverse tim esteps for the prediction of each task. One crucial challenge here is deciding o n the direction and the amount of knowledge transfer, since loss-based knowledge transfer Lee et al. (2016; 2017) does not apply in our case where we do not hav e loss at each timestep. We propose to tackle this challenge by proposing a nove l uncertainty- based probabilistic knowledge transfer mechanism, such that we pe rform knowledge transfer from more certain tasks with lower variance to uncertai n ones with higher variance. We validate our Temporal Probabilistic Asymmetric M ulti-task Learning (TP-AMTL) model on two clinical risk prediction tasks against recent deep learning models for time-series analysis, which our model significa ntly outperforms by successfully preventing negative transfer. Further qualitati ve analysis of our model by clinicians suggests that the learned knowledge trans fer graphs are helpful in analyzing the model's predictions.

****

Lazy-CFR: fast and near-optimal regret minimization for extensive games with imp erfect information

Yichi Zhou,Tongzheng Ren,Jialian Li,Dong Yan,Jun Zhu

Counterfactual regret minimization (CFR) methods are effective for solving two-p layer zero-sum extensive games with imperfect information with state-of-the-art results. However, the vanilla CFR has to traverse the whole game tree in each round, which is time-consuming in large-scale games. In this paper, we present Lazy-CFR, a CFR algorithm that adopts a lazy update strategy to avoid traversing the whole game tree in each round. We prove that the regret of Lazy-CFR is alm ost the same to the regret of the vanilla CFR and only needs to visit a small po rtion of the game tree. Thus, Lazy-CFR is provably faster than CFR. Empirical r esults consistently show that Lazy-CFR is significantly faster than the vanilla CFR.

****

Corpus Based Amharic Sentiment Lexicon Generation

Girma Neshir,Andeas Rauber,and Solomon Atnafu

Sentiment classification is an active research area with several applications in cluding analysis of political opinions, classifying comments, movie reviews, new s reviews and product reviews. To employ rule based sentiment classification, we require sentiment lexicons. However, manual construction of sentiment lexicon i s time consuming and costly for resource-limited languages. To bypass manual dev elopment time and costs, we tried to build Amharic Sentiment Lexicons relying on corpus based approach. The intention of this approach is to handle sentiment te rms specific to Amharic language from Amharic Corpus. Small set of seed terms ar e manually prepared from three parts of speech such as noun, adjective and verb. We developed algorithms for constructing Amharic sentiment lexicons automatical ly from Amharic news corpus. Corpus based approach is proposed relying on the wo rd co-occurrence distributional embedding including frequency based embedding (i .e. Positive Point-wise Mutual Information PPMI). First we build word-context un igram frequency count matrix and transform it to point-wise mutual Information m atrix. Using this matrix, we computed the cosine distance of mean vector of seed lists and each word in the corpus vocabulary. Based on the threshold value, the top closest words to the mean vector of seed list are added to the lexicon. The n the mean vector of the new sentiment seed list is updated and process is repea ted until we get sufficient terms in the lexicon. Using PPMI with threshold valu e of 100 and 200, we got corpus based Amharic Sentiment lexicons of size 1811 an d 3794 respectively by expanding 519 seeds. Finally, the lexicon generated in co rpus based approach is evaluated.

****

Principled Weight Initialization for Hypernetworks

Oscar Chang,Lampros Flokas,Hod Lipson

Hypernetworks are meta neural networks that generate weights for a main neural n etwork in an end-to-end differentiable manner. Despite extensive applications ra nging from multi-task learning to Bayesian deep learning, the problem of optimiz ing hypernetworks has not been studied to date. We observe that classical weight

initialization methods like Glorot & Bengio (2010) and He et al. (2015), when a
pplied directly on a hypernet, fail to produce weights for the mainnet in the co
rrect scale. We develop principled techniques for weight initialization in hyper
nets, and show that they lead to more stable mainnet weights, lower training los
s, and faster convergence.
****
Additive Powers-of-Two Quantization: An Efficient Non-uniform Discretization for
Neural Networks
Yuhang Li,Xin Dong,Wei Wang
We propose Additive Powers-of-Two~(APoT) quantization, an efficient non-uniform
quantization scheme for the bell-shaped and long-tailed distribution of weights
and activations in neural networks. By constraining all quantization levels as t
he sum of Powers-of-Two terms, APoT quantization enjoys high computational effic
iency and a good match with the distribution of weights. A simple reparameteriza
tion of the clipping function is applied to generate a better-defined gradient f
or learning the clipping threshold. Moreover, weight normalization is presented
to refine the distribution of weights to make the training more stable and consi
stent. Experimental results show that our proposed method outperforms state-of-t
he-art methods, and is even competitive with the full-precision models, demonstr
ating the effectiveness of our proposed APoT quantization. For example, our 4-bi
t quantized ResNet-50 on ImageNet achieves 76.6% top-1 accuracy without bells an
d whistles; meanwhile, our model reduces 22% computational cost compared with th
e uniformly quantized counterpart.
****
Transfer Alignment Network for Double Blind Unsupervised Domain Adaptation
Huiwen Xu,U Kang
How can we transfer knowledge from a source domain to a target domain when each
side cannot observe the data in the other side? The recent state-of-the-art deep
architectures show significant performance in classification tasks which highly
depend on a large number of training data. In order to resolve the dearth of ab
undant target labeled data, transfer learning and unsupervised learning leverage
data from different sources and unlabeled data as training data, respectively.
However, in some practical settings, transferring source data to target domain i
s restricted due to a privacy policy.

In this paper, we define the problem of unsupervised domain adaptation under dou
ble blind constraint, where either the source or the target domain cannot observ
e the data in the other domain, but data from both domains are used for training
. We propose TAN (Transfer Alignment Network for Double Blind Domain Adaptation)
, an effective method for the problem by aligning source and target domain featu
res. TAN maps the target feature into source feature space so that the classifie
r learned from the labeled data in the source domain is readily used in the targ
et domain. Extensive experiments show that TAN 1) provides the state-of-the-art
accuracy for double blind domain adaptation, and 2) outperforms baselines regard
less of the proportion of target domain data in the training data.

****
Towards understanding the true loss surface of deep neural networks using random
matrix theory and iterative spectral methods
Diego Granziol,Timur Garipov,Dmitry Vetrov,Stefan Zohren,Stephen Roberts,Andrew
Gordon Wilson
The geometric properties of loss surfaces, such as the local flatness of a solut
ion, are associated with generalization in deep learning. The Hessian is often u
sed to understand these geometric properties. We investigate the differences bet
ween the eigenvalues of the neural network Hessian evaluated over the empirical
dataset, the Empirical Hessian, and the eigenvalues of the Hessian under the dat
a generating distribution, which we term the True Hessian. Under mild assumption
s, we use random matrix theory to show that the True Hessian has eigenvalues of
smaller absolute value than the Empirical Hessian. We support these results for
different SGD schedules on both a 110-Layer ResNet and VGG-16. To perform these

experiments we propose a framework for spectral visualization, based on GPU acce
lerated stochastic Lanczos quadrature. This approach is an order of magnitude fa
ster than state-of-the-art methods for spectral visualization, and can be generi
cally used to investigate the spectral properties of matrices in deep learning.
****

Neural Architecture Search in Embedding Space
chun-ting liu
The neural architecture search (NAS) algorithm with reinforcement learning can b
e a powerful and novel framework for the automatic discovering process of neural
 architectures. However, its application is restricted by noncontinuous and high
-dimensional search spaces, which result in difficulty in optimization. To resol
ve these problems, we proposed NAS in embedding space (NASES), which is a novel
framework. Unlike other NAS with reinforcement learning approaches that search o
ver a discrete and high-dimensional architecture space, this approach enables re
inforcement learning to search in an embedding space by using architecture encod
ers and decoders. The current experiment demonstrated that the performance of th
e final architecture network using the NASES procedure is comparable with that o
f other popular NAS approaches for the image classification task on CIFAR-10. Th
e beneficial-performance and effectiveness of NASES was impressive even when onl
y the architecture-embedding searching and pre-training controller were applied
without other NAS tricks such as parameter sharing. Specifically, considerable r
eduction in searches was achieved by reducing the average number of searching to
 < 100 architectures to achieve a final architecture for the NASES procedure.
****

Enhancing Transformation-Based Defenses Against Adversarial Attacks with a Distr
ibution Classifier
Connie Kou,Hwee Kuan Lee,Ee-Chien Chang,Teck Khim Ng
Adversarial attacks on convolutional neural networks (CNN) have gained significa
nt attention and there have been active research efforts on defense mechanisms.
Stochastic input transformation methods have been proposed, where the idea is to
 recover the image from adversarial attack by random transformation, and to take
 the majority vote as consensus among the random samples. However, the transform
ation improves the accuracy on adversarial images at the expense of the accuracy
 on clean images. While it is intuitive that the accuracy on clean images would
deteriorate, the exact mechanism in which how this occurs is unclear. In this pa
per, we study the distribution of softmax induced by stochastic transformations.
 We observe that with random transformations on the clean images, although the m
ass of the softmax distribution could shift to the wrong class, the resulting di
stribution of softmax could be used to correct the prediction. Furthermore, on t
he adversarial counterparts, with the image transformation, the resulting shapes
 of the distribution of softmax are similar to the distributions from the clean
images. With these observations, we propose a method to improve existing transfo
rmation-based defenses. We train a separate lightweight distribution classifier
to recognize distinct features in the distributions of softmax outputs of transf
ormed images. Our empirical studies show that our distribution classifier, by tr
aining on distributions obtained from clean images only, outperforms majority vo
ting for both clean and adversarial images. Our method is generic and can be int
egrated with existing transformation-based defenses.
****

HaarPooling: Graph Pooling with Compressive Haar Basis
Yu Guang Wang,Ming Li,Zheng Ma,Guido Montufar,Xiaosheng Zhuang,Yanan Fan
Deep Graph Neural Networks (GNNs) are instrumental in graph classification and g
raph-based regression tasks. In these tasks, graph pooling is a critical ingredi
ent by which GNNs adapt to input graphs of varying size and structure. We propos
e a new graph pooling operation based on compressive Haar transforms, called Haa
rPooling. HaarPooling is computed following a chain of sequential clusterings of
 the input graph. The input of each pooling layer is transformed by the compress
ive Haar basis of the corresponding clustering. HaarPooling operates in the freq
uency domain by the synthesis of nodes in the same cluster and filters out fine
detail information by compressive Haar transforms. Such transforms provide an ef

fective characterization of the data and preserve the structure information of t
he input graph. By the sparsity of the Haar basis, the computation of HaarPoolin
g is of linear complexity. The GNN with HaarPooling and existing graph convoluti
on layers achieves state-of-the-art performance on diverse graph classification
problems.
****
Safe Policy Learning for Continuous Control
Yinlam Chow,Ofir Nachum,Aleksandra Faust,Edgar Duenez-Guzman,Mohammad Ghavamzade
h
We study continuous action reinforcement learning problems in which it is crucia
l that the agent interacts with the environment only through safe policies, i.e.
,~policies that keep the agent in desirable situations, both during training and
 at convergence. We formulate these problems as {\em constrained} Markov decisio
n processes (CMDPs) and present safe policy optimization algorithms that are bas
ed on a Lyapunov approach to solve them. Our algorithms can use any standard pol
icy gradient (PG) method, such as deep deterministic policy gradient (DDPG) or p
roximal policy optimization (PPO), to train a neural network policy, while guara
nteeing near-constraint satisfaction for every policy update by projecting eithe
r the policy parameter or the selected action onto the set of feasible solutions
 induced by the state-dependent linearized Lyapunov constraints. Compared to the
 existing constrained PG algorithms, ours are more data efficient as they are ab
le to utilize both on-policy and off-policy data. Moreover, our action-projectio
n algorithm often leads to less conservative policy updates and allows for natur
al integration into an end-to-end PG training pipeline. We evaluate our algorith
ms and compare them with the state-of-the-art baselines on several simulated (Mu
JoCo) tasks, as well as a real-world robot obstacle-avoidance problem, demonstra
ting their effectiveness in terms of balancing performance and constraint satisf
action.
****
A Stochastic Trust Region Method for Non-convex Minimization
Zebang Shen,Pan Zhou,Cong Fang,Jiahao Xie,Alejandro Ribeiro
We target the problem of finding a local minimum in non-convex finite-sum minimi
zation. Towards this goal, we first prove that the trust region method with inex
act gradient and Hessian estimation can achieve a convergence rate of order $\ma
thcal{O}({1}/{k^{2/3}})$ as long as those differential estimations are sufficien
tly accurate.
Combining such result with a novel Hessian estimator, we propose a sample-effici
ent stochastic trust region (STR) algorithm which finds an $(\epsilon, \sqrt{\ep
silon})$-approximate local minimum within $\tilde{\mathcal{O}}({\sqrt{n}}/{\epsi
lon^{1.5}})$ stochastic Hessian oracle queries.
This improves the state-of-the-art result by a factor of $\mathcal{O}(n^{1/6})$.
 Finally, we also develop Hessian-free STR algorithms which achieve the lowest
runtime complexity.
Experiments verify theoretical conclusions and the efficiency of the proposed al
gorithms.
****
Learning Effective Exploration Strategies For Contextual Bandits
Amr Sharaf,Hal Daumé III
In contextual bandits, an algorithm must choose actions given observed contexts,
 learning from a reward signal that is observed only for the action chosen. This
 leads to an exploration/exploitation trade-off: the algorithm must balance taki
ng actions it already believes are good with taking new actions to potentially d
iscover better choices. We develop a meta-learning algorithm, MELEE, that learns
 an exploration policy based on simulated, synthetic contextual bandit tasks. ME
LEE uses imitation learning against these simulations to train an exploration po
licy that can be applied to true contextual bandit tasks at test time. We evalua
te on both a natural contextual bandit problem derived from a learning to rank d
ataset as well as hundreds of simulated contextual bandit problems derived from
classification tasks. MELEE outperforms seven strong baselines on most of these
datasets by leveraging a rich feature representation for learning an exploration

strategy.
****
Improving Batch Normalization with Skewness Reduction for Deep Neural Networks
Pak Lun Kevin Ding,Sarah Martin,Baoxin Li
Batch Normalization (BN) is a well-known technique used in training deep neural networks.
The main idea behind batch normalization is to normalize the features of the layers ($i.e.$, transforming them to have a mean equal to zero and a variance equal to one).
Such a procedure encourages the optimization landscape of the loss function to be smoother, and improve the learning of the networks for both speed and performance.
In this paper,
we demonstrate that the performance of the network can be improved,
if the distributions of the features of the output in the same layer are similar.
As normalizing based on mean and variance does not necessarily make the features to have the same distribution, we propose a new normalization scheme: Batch Normalization with Skewness Reduction (BNSR).
Comparing with other normalization approaches,
BNSR transforms not just only the mean and variance,
but also the skewness of the data.
By tackling this property of a distribution, we are able to make the output distributions of the layers to be further similar. The nonlinearity of BNSR may further improve the expressiveness of the underlying network.
Comparisons with other normalization schemes are tested on the CIFAR-100 and ImageNet datasets. Experimental results show that the proposed approach can outperform other state-of-the-arts that are not equipped with BNSR.
****
Adversarial Inductive Transfer Learning with input and output space adaptation
Hossein Sharifi-Noghabi,Shuman Peng,Olga Zolotareva,Colin C. Collins,Martin Ester
We propose Adversarial Inductive Transfer Learning (AITL), a method for addressing discrepancies in input and output spaces between source and target domains. AITL utilizes adversarial domain adaptation and multi-task learning to address these discrepancies. Our motivating application is pharmacogenomics where the goal is to predict drug response in patients using their genomic information. The challenge is that clinical data (i.e. patients) with drug response outcome is very limited, creating a need for transfer learning to bridge the gap between large pre-clinical pharmacogenomics datasets (e.g. cancer cell lines) and clinical datasets. Discrepancies exist between 1) the genomic data of pre-clinical and clinical datasets (the input space), and 2) the different measures of the drug response (the output space). To the best of our knowledge, AITL is the first adversarial inductive transfer learning method to address both input and output discrepancies. Experimental results indicate that AITL outperforms state-of-the-art pharmacogenomics and transfer learning baselines and may guide precision oncology more accurately.
****
Graph Neural Networks For Multi-Image Matching
Stephen Phillips,Kostas Daniilidis
In geometric computer vision applications, multi-image feature matching gives more accurate and robust solutions compared to simple two-image matching. In this work, we formulate multi-image matching as a graph embedding problem, then use a Graph Neural Network to learn an appropriate embedding function for aligning image features. We use cycle consistency to train our network in an unsupervised fashion, since ground truth correspondence can be difficult or expensive to acquire. Geometric consistency losses are added to aid training, though unlike optimization based methods no geometric information is necessary at inference time. To the best of our knowledge, no other works have used graph neural networks for multi-image feature matching. Our experiments show that our method is competitive

with other optimization based approaches.
****
An Empirical Study on Post-processing Methods for Word Embeddings
Shuai Tang,Mahta Mousavi,Virginia R. de Sa
Word embeddings learnt from large corpora have been adopted in various applications in natural language processing and served as the general input representations to learning systems. Recently, a series of post-processing methods have been proposed to boost the performance of word embeddings on similarity comparison and analogy retrieval tasks, and some have been adapted to compose sentence representations. The general hypothesis behind these methods is that by enforcing the embedding space to be more isotropic, the similarity between words can be better expressed. We view these methods as an approach to shrink the covariance/gram matrix, which is estimated by learning word vectors, towards a scaled identity matrix. By optimising an objective in the semi-Riemannian manifold with Centralised Kernel Alignment (CKA), we are able to search for the optimal shrinkage parameter, and provide a post-processing method to smooth the spectrum of learnt word vectors which yields improved performance on downstream tasks.
****
AN EFFICIENT HOMOTOPY TRAINING ALGORITHM FOR NEURAL NETWORKS
Qipin Chen,Wenrui Hao
We present a Homotopy Training Algorithm (HTA) to solve optimization problems arising from neural networks. The HTA starts with several decoupled systems with low dimensional structure and tracks the solution to the high dimensional coupled system. The decoupled systems are easy to solve due to the low dimensionality but can be connected to the original system via a continuous homotopy path guided by the HTA. We have proved the convergence of HTA for the non-convex case and existence of the homotopy solution path for the convex case. The HTA has provided a better accuracy on several examples including VGG models on CIFAR-10. Moreover, the HTA would be combined with the dropout technique to provide an alternative way to train the neural networks.
****
High performance RNNs with spiking neurons
Manu V Nair,Giacomo Indiveri
The increasing need for compact and low-power computing solutions for machine learning applications has triggered a renaissance in the study of energy-efficient neural network accelerators. In particular, in-memory computing neuromorphic architectures have started to receive substantial attention from both academia and industry. However, most of these architectures rely on spiking neural networks, which typically perform poorly compared to their non-spiking counterparts in terms of accuracy. In this paper, we propose a new adaptive spiking neuron model that can also be abstracted as a low-pass filter. This abstraction enables faster and better training of spiking networks using back-propagation, without simulating spikes. We show that this model dramatically improves the inference performance of a recurrent neural network and validate it with three complex spatio-temporal learning tasks: the temporal addition task, the temporal copying task, and a spoken-phrase recognition task. Application of these results will lead to the development of powerful spiking models for neuromorphic hardware that solve relevant edge-computing and Internet-of-Things applications with high accuracy and ultra-low power consumption.
****
CLAREL: classification via retrieval loss for zero-shot learning
Boris N. Oreshkin,Negar Rostamzadeh,Pedro O. Pinheiro,Christopher Pal
We address the problem of learning fine-grained cross-modal representations. We propose an instance-based deep metric learning approach in joint visual and textual space. The key novelty of this paper is that it shows that using per-image semantic supervision leads to substantial improvement in zero-shot performance over using class-only supervision. On top of that, we provide a probabilistic justification for a metric rescaling approach that solves a very common problem in the generalized zero-shot learning setting, i.e., classifying test images from unseen classes as one of the classes seen during training. We evaluate our approac

h on two fine-grained zero-shot learning datasets: CUB and FLOWERS. We find that on the generalized zero-shot classification task CLAREL consistently outperforms the existing approaches on both datasets.
****

Observational Overfitting in Reinforcement Learning
Xingyou Song,Yiding Jiang,Stephen Tu,Yilun Du,Behnam Neyshabur
A major component of overfitting in model-free reinforcement learning (RL) involves the case where the agent may mistakenly correlate reward with certain spurious features from the observations generated by the Markov Decision Process (MDP). We provide a general framework for analyzing this scenario, which we use to design multiple synthetic benchmarks from only modifying the observation space of an MDP. When an agent overfits to different observation spaces even if the underlying MDP dynamics is fixed, we term this observational overfitting. Our experiments expose intriguing properties especially with regards to implicit regularization, and also corroborate results from previous works in RL generalization and supervised learning (SL).
****

On Mutual Information Maximization for Representation Learning
Michael Tschannen,Josip Djolonga,Paul K. Rubenstein,Sylvain Gelly,Mario Lucic
Many recent methods for unsupervised or self-supervised representation learning train feature extractors by maximizing an estimate of the mutual information (MI) between different views of the data. This comes with several immediate problems: For example, MI is notoriously hard to estimate, and using it as an objective for representation learning may lead to highly entangled representations due to its invariance under arbitrary invertible transformations. Nevertheless, these methods have been repeatedly shown to excel in practice. In this paper we argue, and provide empirical evidence, that the success of these methods cannot be attributed to the properties of MI alone, and that they strongly depend on the inductive bias in both the choice of feature extractor architectures and the parametrization of the employed MI estimators. Finally, we establish a connection to deep metric learning and argue that this interpretation may be a plausible explanation for the success of the recently introduced methods.
****

Localizing and Amortizing: Efficient Inference for Gaussian Processes
Linfeng Liu,Liping Liu
The inference of Gaussian Processes concerns the distribution of the underlying function given observed data points. GP inference based on local ranges of data points is able to capture fine-scale correlations and allow fine-grained decomposition of the computation. Following this direction, we propose a new inference model that considers the correlations and observations of the K nearest neighbors for the inference at a data point. Compared with previous works, we also eliminate the data ordering prerequisite to simplify the inference process. Additionally, the inference task is decomposed to small subtasks with several technique innovations, making our model well suits the stochastic optimization. Since the decomposed small subtasks have the same structure, we further speed up the inference procedure with amortized inference. Our model runs efficiently and achieves good performances on several benchmark tasks.
****

PNAT: Non-autoregressive Transformer by Position Learning
Yu Bao,Hao Zhou,Jiangtao Feng,Mingxuan Wang,Shujian Huang,Jiajun Chen,Lei Li
Non-autoregressive generation is a new paradigm for text generation. Previous work hardly considers to explicitly model the positions of generated words. However, position modeling of output words is an essential problem in non-autoregressive text generation. In this paper, we propose PNAT, which explicitly models positions of output words as latent variables in text generation. The proposed PNATis simple yet effective. Experimental results show that PNATgives very promising results in machine translation and paraphrase generation tasks, outperforming many strong baselines.
****

On unsupervised-supervised risk and one-class neural networks

Christophe Cerisara
Most unsupervised neural networks training methods concern generative models, de
ep clustering, pretraining or some form of representation learning. We rather de
al in this work with unsupervised training of the final classification stage of
a standard deep learning stack, with a focus on two types of methods: unsupervis
ed-supervised risk approximations and one-class models. We derive a new analytic
al solution for the former and identify and analyze its similarity with the latt
er.
We apply and validate the proposed approach on multiple experimental conditions,
 in particular on four challenging recent Natural Language Processing tasks as w
ell as on an anomaly detection task, where it improves over state-of-the-art mod
els.
****
Tranquil Clouds: Neural Networks for Learning Temporally Coherent Features in Po
int Clouds
Lukas Prantl,Nuttapong Chentanez,Stefan Jeschke,Nils Thuerey
Point clouds, as a form of Lagrangian representation, allow for powerful and fle
xible applications in a large number of computational disciplines. We propose a
novel deep-learning method to learn stable and temporally coherent feature space
s for points clouds that change over time. We identify a set of inherent problem
s with these approaches: without knowledge of the time dimension, the inferred s
olutions can exhibit strong flickering, and easy solutions to suppress this flic
kering can result in undesirable local minima that manifest themselves as halo s
tructures. We propose a novel temporal loss function that takes into account hig
her time derivatives of the point positions, and encourages mingling, i.e., to p
revent the aforementioned halos. We combine these techniques in a super-resoluti
on method with a truncation approach to flexibly adapt the size of the generated
 positions. We show that our method works for large, deforming point sets from d
ifferent sources to demonstrate the flexibility of our approach.
****
Distillation $\approx$ Early Stopping? Harvesting Dark Knowledge Utilizing Aniso
tropic Information Retrieval For Overparameterized NN
Bin Dong,Jikai Hou,Yiping Lu,Zhihua Zhang
Distillation is a method to transfer knowledge from one model to another and oft
en achieves higher accuracy with the same capacity. In this paper, we aim to pro
vide a theoretical understanding on what mainly helps with the distillation. Our
 answer is "early stopping". Assuming that the teacher network is overparameteri
zed, we argue that the teacher network is essentially harvesting dark knowledge
from the data via early stopping. This can be justified by a new concept, Anisot
ropic In- formation Retrieval (AIR), which means that the neural network tends t
o fit the informative information first and the non-informative information (inc
luding noise) later. Motivated by the recent development on theoretically analyz
ing overparame- terized neural networks, we can characterize AIR by the eigenspa
ce of the Neural Tangent Kernel(NTK). AIR facilities a new understanding of dist
illation. With that, we further utilize distillation to refine noisy labels. We
propose a self-distillation al- gorithm to sequentially distill knowledge from t
he network in the previous training epoch to avoid memorizing the wrong labels.
We also demonstrate, both theoret- ically and empirically, that self-distillatio
n can benefit from more than just early stopping. Theoretically, we prove conver
gence of the proposed algorithm to the ground truth labels for randomly initiali
zed overparameterized neural networks in terms of l2 distance, while the previou
s result was on convergence in 0-1 loss. The theoretical result ensures the lear
ned neural network enjoy a margin on the training data which leads to better gen
eralization. Empirically, we achieve better testing accuracy and entirely avoid
early stopping which makes the algorithm more user-friendly.

****
Bayesian Inference for Large Scale Image Classification
Jonathan Heek,Nal Kalchbrenner
Bayesian inference promises to ground and improve the performance of deep neural

networks. It promises to be robust to overfitting, to simplify the training procedure and the space of hyperparameters, and to provide a calibrated measure of uncertainty that can enhance decision making, agent exploration and prediction fairness.
Markov Chain Monte Carlo (MCMC) methods enable Bayesian inference by generating samples from the posterior distribution over model parameters.
Despite the theoretical advantages of Bayesian inference and the similarity between MCMC and optimization methods, the performance of sampling methods has so far lagged behind optimization methods for large scale deep learning tasks.
We aim to fill this gap and introduce ATMC, an adaptive noise MCMC algorithm that estimates and is able to sample from the posterior of a neural network.
ATMC dynamically adjusts the amount of momentum and noise applied to each parameter update in order to compensate for the use of stochastic gradients.
We use a ResNet architecture without batch normalization to test ATMC on the Cifar10 benchmark and the large scale ImageNet benchmark and show that, despite the absence of batch normalization, ATMC outperforms a strong optimization baseline in terms of both classification accuracy and test log-likelihood. We show that ATMC is intrinsically robust to overfitting on the training data and that ATMC provides a better calibrated measure of uncertainty compared to the optimization baseline.
****

Ranking Policy Gradient
Kaixiang Lin,Jiayu Zhou
Sample inefficiency is a long-lasting problem in reinforcement learning (RL). The state-of-the-art estimates the optimal action values while it usually involves an extensive search over the state-action space and unstable optimization. Towards the sample-efficient RL, we propose ranking policy gradient (RPG), a policy gradient method that learns the optimal rank of a set of discrete actions. To accelerate the learning of policy gradient methods, we establish the equivalence between maximizing the lower bound of return and imitating a near-optimal policy without accessing any oracles. These results lead to a general off-policy learning framework, which preserves the optimality, reduces variance, and improves the sample-efficiency. We conduct extensive experiments showing that when consolidating with the off-policy learning framework, RPG substantially reduces the sample complexity, comparing to the state-of-the-art.
****

How Does Learning Rate Decay Help Modern Neural Networks?
Kaichao You,Mingsheng Long,Jianmin Wang,Michael I. Jordan
Learning rate decay (lrDecay) is a \emph{de facto} technique for training modern neural networks. It starts with a large learning rate and then decays it multiple times. It is empirically observed to help both optimization and generalization. Common beliefs in how lrDecay works come from the optimization analysis of (Stochastic) Gradient Descent: 1) an initially large learning rate accelerates training or helps the network escape spurious local minima; 2) decaying the learning rate helps the network converge to a local minimum and avoid oscillation. Despite the popularity of these common beliefs, experiments suggest that they are insufficient in explaining the general effectiveness of lrDecay in training modern neural networks that are deep, wide, and nonconvex. We provide another novel explanation: an initially large learning rate suppresses the network from memorizing noisy data while decaying the learning rate improves the learning of complex patterns. The proposed explanation is validated on a carefully-constructed dataset with tractable pattern complexity. And its implication, that additional patterns learned in later stages of lrDecay are more complex and thus less transferable, is justified in real-world datasets. We believe that this alternative explanation will shed light into the design of better training strategies for modern neural networks.
****

Random Matrix Theory Proves that Deep Learning Representations of GAN-data Behave as Gaussian Mixtures
Mohamed El Amine Seddik,Cosme Louart,Mohamed Tamaazousti,Romain Couillet

This paper shows that deep learning (DL) representations of data produced by gen erative adversarial nets (GANs) are random vectors which fall within the class o f so-called concentrated random vectors. Further exploiting the fact that Gram m atrices, of the type G = X'X with X = [x_1 , . . . , x_n] ∈ R p×n and x_i indep endent concentrated random vectors from a mixture model, behave asymptotically (as n, p → ∞) as if the x_i were drawn from a Gaussian mixture, suggests that DL representations of GAN-data can be fully described by their first two statistica l moments for a wide range of standard classifiers. Our theoretical findings are validated by generating images with the BigGAN model and across different popul ar deep representation networks.
****

SVQN: Sequential Variational Soft Q-Learning Networks
Shiyu Huang,Hang Su,Jun Zhu,Ting Chen
Partially Observable Markov Decision Processes (POMDPs) are popular and flexible models for real-world decision-making applications that demand the information from past observations to make optimal decisions. Standard reinforcement learnin g algorithms for solving Markov Decision Processes (MDP) tasks are not applicabl e, as they cannot infer the unobserved states. In this paper, we propose a novel algorithm for POMDPs, named sequential variational soft Q-learning networks (SV QNs), which formalizes the inference of hidden states and maximum entropy reinfo rcement learning (MERL) under a unified graphical model and optimizes the two mo dules jointly. We further design a deep recurrent neural network to reduce the c omputational complexity of the algorithm. Experimental results show that SVQNs c an utilize past information to help decision making for efficient inference, and outperforms other baselines on several challenging tasks. Our ablation study sh ows that SVQNs have the generalization ability over time and are robust to the d isturbance of the observation.
****

Classification Attention for Chinese NER
Yuchen Ge,FanYang,PeiYang
The character-based model, such as BERT, has achieved remarkable success in Chin ese named entity recognition (NER). However, such model would likely miss the ov erall information of the entity words. In this paper, we propose to combine prio ri entity information with BERT. Instead of relying on additional lexicons or pr e-trained word embeddings, our model has generated entity classification embeddi ngs directly on the pre-trained BERT, having the merit of increasing model pract icability and avoiding OOV problem. Experiments show that our model has achieved state-of-the-art results on 3 Chinese NER datasets.
****

Understanding Isomorphism Bias in Graph Data Sets
Ivanov Sergey,Sviridov Sergey,Evgeny Burnaev
In recent years there has been a rapid increase in classification methods on gra ph structured data. Both in graph kernels and graph neural networks, one of the implicit assumptions of successful state-of-the-art models was that incorporatin g graph isomorphism features into the architecture leads to better empirical per formance. However, as we discover in this work, commonly used data sets for grap h classification have repeating instances which cause the problem of isomorphism bias, i.e. artificially increasing the accuracy of the models by memorizing tar get information from the training set. This prevents fair competition of the alg orithms and raises a question of the validity of the obtained results. We analyz e 54 data sets, previously extensively used for graph-related tasks, on the exis tence of isomorphism bias, give a set of recommendations to machine learning pra ctitioners to properly set up their models, and open source new data sets for th e future experiments.
****

Neural Machine Translation with Universal Visual Representation
Zhuosheng Zhang,Kehai Chen,Rui Wang,Masao Utiyama,Eiichiro Sumita,Zuchao Li,Hai Zhao
Though visual information has been introduced for enhancing neural machine trans lation (NMT), its effectiveness strongly relies on the availability of large amo

unts of bilingual parallel sentence pairs with manual image annotations. In this paper, we present a universal visual representation learned over the monolingual corpora with image annotations, which overcomes the lack of large-scale bilingual sentence-image pairs, thereby extending image applicability in NMT. In detail, a group of images with similar topics to the source sentence will be retrieved from a light topic-image lookup table learned over the existing sentence-image pairs, and then is encoded as image representations by a pre-trained ResNet. An attention layer with a gated weighting is to fuse the visual information and text information as input to the decoder for predicting target translations. In particular, the proposed method enables the visual information to be integrated into large-scale text-only NMT in addition to the multimodel NMT. Experiments on four widely used translation datasets, including the WMT'16 English-to-Romanian, WMT'14 English-to-German, WMT'14 English-to-French, and Multi30K, show that the proposed approach achieves significant improvements over strong baselines.
****

Towards More Realistic Neural Network Uncertainties
Joachim Sicking,Alexander Kister,Matthias Fahrland,Stefan Eickeler,Fabian Hueger,Stefan Rueping,Peter Schlicht,Tim Wirtz
Statistical models are inherently uncertain. Quantifying or at least upper-bounding their uncertainties is vital for safety-critical systems. While standard neural networks do not report this information, several approaches exist to integrate uncertainty estimates into them. Assessing the quality of these uncertainty estimates is not straightforward, as no direct ground truth labels are available. Instead, implicit statistical assessments are required. For regression, we propose to evaluate uncertainty realism---a strict quality criterion---with a Mahalanobis distance-based statistical test. An empirical evaluation reveals the need for uncertainty measures that are appropriate to upper-bound heavy-tailed empirical errors. Alongside, we transfer the variational U-Net classification architecture to standard supervised image-to-image tasks. It provides two uncertainty mechanisms and significantly improves uncertainty realism compared to a plain encoder-decoder model.
****

Understanding Architectures Learnt by Cell-based Neural Architecture Search
Yao Shu,Wei Wang,Shaofeng Cai
Neural architecture search (NAS) searches architectures automatically for given tasks, e.g., image classification and language modeling. Improving the search efficiency and effectiveness has attracted increasing attention in recent years. However, few efforts have been devoted to understanding the generated architectures. In this paper, we first reveal that existing NAS algorithms (e.g., DARTS, ENAS) tend to favor architectures with wide and shallow cell structures. These favorable architectures consistently achieve fast convergence and are consequently selected by NAS algorithms. Our empirical and theoretical study further confirms that their fast convergence derives from their smooth loss landscape and accurate gradient information. Nonetheless, these architectures may not necessarily lead to better generalization performance compared with other candidate architectures in the same search space, and therefore further improvement is possible by revising existing NAS algorithms.
****

Soft Token Matching for Interpretable Low-Resource Classification
Federico Errica,Fabrizio Silvestri,Bora Edizel,Sebastian Riedel,Ludovic Denoyer,Vassilis Plachouras
We propose a model to tackle classification tasks in the presence of very little training data. To this aim, we introduce a novel matching mechanism to focus on elements of the input by using vectors that represent semantically meaningful concepts for the task at hand.
By leveraging highlighted portions of the training data, a simple, yet effective, error boosting technique guides the learning process. In practice, it increases the error associated to relevant parts of the input by a given factor. Results on text classification tasks confirm the benefits of the proposed approach in both balanced and unbalanced cases, thus being of practical use when labeling new

examples is expensive. In addition, the model is interpretable, as it allows fo
r human inspection of the learned weights.
****
Beyond Classical Diffusion: Ballistic Graph Neural Network
Yimeng Min
This paper presents the ballistic graph neural network. Ballistic graph neural n
etwork tackles the weight distribution from a transportation perspective and has
many different properties comparing to the traditional graph neural network pip
eline. The ballistic graph neural network does not require to calculate any eige
nvalue. The filters propagate exponentially faster($\sigma^2 \sim T^2$) comparin
g to traditional graph neural network($\sigma^2 \sim T$). We use a perturbed coi
n operator to perturb and optimize the diffusion rate. Our results show that by
selecting the diffusion speed, the network can reach a similar accuracy with few
er parameters. We also show the perturbed filters act as better representations
comparing to pure ballistic ones. We provide a new perspective of training graph
neural network, by adjusting the diffusion rate, the neural network's performan
ce can be improved.
****
Understanding and Stabilizing GANs' Training Dynamics with Control Theory
Kun Xu,Chongxuan Li,Huanshu Wei,Jun Zhu,Bo Zhang
Generative adversarial networks~(GANs) have made significant progress on realist
ic image generation but often suffer from instability during the training proces
s. Most previous analyses mainly focus on the equilibrium that GANs achieve, whe
reas a gap exists between such theoretical analyses and practical implementation
s, where it is the training dynamics that plays a vital role in the convergence
and stability of GANs. In this paper, we directly model the dynamics of GANs and
adopt the control theory to understand and stabilize it. Specifically, we inter
pret the training process of various GANs as certain types of dynamics in a unif
ied perspective of control theory which enables us to model the stability and co
nvergence easily. Borrowed from control theory, we adopt the widely-used negativ
e feedback control to stabilize the training dynamics, which can be considered a
s an $L2$ regularization on the output of the discriminator. We empirically veri
fy our method on both synthetic data and natural image datasets. The results dem
onstrate that our method can stabilize the training dynamics as well as converge
better than baselines.
****
Variance Reduced Local SGD with Lower Communication Complexity
Xianfeng Liang,Shuheng Shen,Jingchang Liu,Zhen Pan,Yifei Cheng,Enhong Chen
To accelerate the training of machine learning models, distributed stochastic gr
adient descent (SGD) and its variants have been widely adopted, which apply mult
iple workers in parallel to speed up training. Among them, Local SGD has gained
much attention due to its lower communication cost. Nevertheless, when the data
distribution on workers is non-identical, Local SGD requires $O(T^{\frac{3}{4}}
N^{\frac{3}{4}})$ communications to maintain its \emph{linear iteration speedup}
property, where T is the total number of iterations and N is the number of
workers. In this paper, we propose Variance Reduced Local SGD (VRL-SGD) to furth
er reduce the communication complexity. Benefiting from eliminating the dependen
cy on the gradient variance among workers, we theoretically prove that VRL-SGD a
chieves a \emph{linear iteration speedup} with a lower communication complexity
$O(T^{\frac{1}{2}} N^{\frac{3}{2}})$ even if workers access non-identical datase
ts. We conduct experiments on three machine learning tasks, and the experimental
results demonstrate that VRL-SGD performs impressively better than Local SGD wh
en the data among workers are quite diverse.
****
AutoQ: Automated Kernel-Wise Neural Network Quantization
Qian Lou,Feng Guo,Minje Kim,Lantao Liu,Lei Jiang.
Network quantization is one of the most hardware friendly techniques to enable t
he deployment of convolutional neural networks (CNNs) on low-power mobile device
s. Recent network quantization techniques quantize each weight kernel in a convo
lutional layer independently for higher inference accuracy, since the weight ker

nels in a layer exhibit different variances and hence have different amounts of redundancy. The quantization bitwidth or bit number (QBN) directly decides the inference accuracy, latency, energy and hardware overhead. To effectively reduce the redundancy and accelerate CNN inferences, various weight kernels should be quantized with different QBNs. However, prior works use only one QBN to quantize each convolutional layer or the entire CNN, because the design space of searching a QBN for each weight kernel is too large. The hand-crafted heuristic of the kernel-wise QBN search is so sophisticated that domain experts can obtain only sub-optimal results. It is difficult for even deep reinforcement learning (DRL) DDPG-based agents to find a kernel-wise QBN configuration that can achieve reasonable inference accuracy. In this paper, we propose a hierarchical-DRL-based kernel-wise network quantization technique, AutoQ, to automatically search a QBN for each weight kernel, and choose another QBN for each activation layer. Compared to the models quantized by the state-of-the-art DRL-based schemes, on average, the same models quantized by AutoQ reduce the inference latency by 54.06%, and decrease the inference energy consumption by 50.69%, while achieving the same inference accuracy.

****

GDP: Generalized Device Placement for Dataflow Graphs

Yanqi Zhou,Sudip Roy,Amirali Abdolrashidi,Daniel Wong,Peter C. Ma,Qiumin Xu,Ming Zhong,Hanxiao Liu,Anna Goldie,Azalia Mirhoseini,James Laudon

Runtime and scalability of large neural networks can be significantly affected by the placement of operations in their dataflow graphs on suitable devices. With increasingly complex neural network architectures and heterogeneous device characteristics, finding a reasonable placement is extremely challenging even for domain experts. Most existing automated device placement approaches are impractical due to the significant amount of compute required and their inability to generalize to new, previously held-out graphs. To address both limitations, we propose an efficient end-to-end method based on a scalable sequential attention mechanism over a graph neural network that is transferable to new graphs. On a diverse set of representative deep learning models, including Inception-v3, AmoebaNet, Transformer-XL, and WaveNet, our method on average achieves 16% improvement over human experts and 9.2% improvement over the prior art with 15 times faster convergence. To further reduce the computation cost, we pre-train the policy network on a set of dataflow graphs and use a superposition network to fine-tune it on each individual graph, achieving state-of-the-art performance on large hold-out graphs with over 50k nodes, such as an 8-layer GNMT.

****

White Noise Analysis of Neural Networks

Ali Borji,Sikun Lin

A white noise analysis of modern deep neural networks is presented to unveil their biases at the whole network level or the single neuron level. Our analysis is
based on two popular and related methods in psychophysics and neurophysiology namely classification images and spike triggered analysis. These methods have been widely used to understand the underlying mechanisms of sensory systems in humans and monkeys. We leverage them to investigate the inherent biases of deep neural networks and to obtain a first-order approximation of their functionality.
We emphasize on CNNs since they are currently the state of the art methods in computer vision and are a decent model of human visual processing. In addition, we study multi-layer perceptrons, logistic regression, and recurrent neural
networks. Experiments over four classic datasets, MNIST, Fashion-MNIST, CIFAR-10, and ImageNet, show that the computed bias maps resemble the target classes and when used for classification lead to an over two-fold performance than
the chance level. Further, we show that classification images can be used to attack
a black-box classifier and to detect adversarial patch attacks. Finally, we util

ize
spike triggered averaging to derive the filters of CNNs and explore how the beha
vior
of a network changes when neurons in different layers are modulated. Our
effort illustrates a successful example of borrowing from neurosciences to study
ANNs and highlights the importance of cross-fertilization and synergy across mac
hine
learning, deep learning, and computational neuroscience.
****

Why Learning of Large-Scale Neural Networks Behaves Like Convex Optimization
Hui Jiang
In this paper, we present some theoretical work to explain why simple gradient d
escent methods are so successful in solving non-convex optimization problems in
learning large-scale neural networks (NN). After introducing a mathematical tool
 called canonical space, we have proved that the objective functions in learnin
g NNs are convex in the canonical model space. We further elucidate that the gra
dients between the original NN model space and the canonical space are related b
y a pointwise linear transformation, which is represented by the so-called disp
arity matrix. Furthermore, we have proved that gradient descent methods surely c
onverge to a global minimum of zero loss provided that the disparity matrices ma
intain full rank. If this full-rank condition holds, the learning of NNs behave
s in the same way as normal convex optimization. At last, we have shown that the
 chance to have singular disparity matrices is extremely slim in large NNs. In p
articular, when over-parameterized NNs are randomly initialized, the gradient de
cent algorithms converge to a global minimum of zero loss in probability.
****

Poly-encoders: Architectures and Pre-training Strategies for Fast and Accurate M
ulti-sentence Scoring
Samuel Humeau,Kurt Shuster,Marie-Anne Lachaux,Jason Weston
The use of deep pre-trained transformers has led to remarkable progress in a num
ber of applications (Devlin et al., 2018). For tasks that make pairwise comparis
ons between sequences, matching a given input with a corresponding label, two ap
proaches are common: Cross-encoders performing full self-attention over the pair
 and Bi-encoders encoding the pair separately. The former often performs better,
 but is too slow for practical use. In this work, we develop a new transformer a
rchitecture, the Poly-encoder, that learns global rather than token level self-a
ttention features. We perform a detailed comparison of all three approaches, inc
luding what pre-training and fine-tuning strategies work best. We show our model
s achieve state-of-the-art results on four tasks; that Poly-encoders are faster
than Cross-encoders and more accurate than Bi-encoders; and that the best result
s are obtained by pre-training on large datasets similar to the downstream tasks
.
****

HighRes-net: Multi-Frame Super-Resolution by Recursive Fusion
Michel Deudon,Alfredo Kalaitzis,Md Rifat Arefin,Israel Goytom,Zhichao Lin,Kris S
ankaran,Vincent Michalski,Samira E Kahou,Julien Cornebise,Yoshua Bengio
Generative deep learning has sparked a new wave of Super-Resolution (SR) algorit
hms that enhance single images with impressive aesthetic results, albeit with im
aginary details. Multi-frame Super-Resolution (MFSR) offers a more grounded appr
oach to the ill-posed problem, by conditioning on multiple low-resolution views.
 This is important for satellite monitoring of human impact on the planet -- fro
m deforestation, to human rights violations -- that depend on reliable imagery.
To this end, we present HighRes-net, the first deep learning approach to MFSR th
at learns its sub-tasks in an end-to-end fashion: (i) co-registration, (ii) fusi
on, (iii) up-sampling, and (iv) registration-at-the-loss. Co-registration of low
-res views is learned implicitly through a reference-frame channel, with no expl
icit registration mechanism. We learn a global fusion operator that is applied r
ecursively on an arbitrary number of low-res pairs. We introduce a registered lo
ss, by learning to align the SR output to a ground-truth through ShiftNet. We sh
ow that by learning deep representations of multiple views, we can super-resolve

low-resolution signals and enhance Earth observation data at scale. Our approach recently topped the European Space Agency's MFSR competition on real-world satellite imagery.
****

A Learning-based Iterative Method for Solving Vehicle Routing Problems
Hao Lu,Xingwen Zhang,Shuang Yang
This paper is concerned with solving combinatorial optimization problems, in particular, the capacitated vehicle routing problems (CVRP). Classical Operations Research (OR) algorithms such as LKH3 \citep{helsgaun2017extension} are inefficient and difficult to scale to larger-size problems. Machine learning based approaches have recently shown to be promising, partly because of their efficiency (once trained, they can perform solving within minutes or even seconds). However, there is still a considerable gap between the quality of a machine learned solution and what OR methods can offer (e.g., on CVRP-100, the best result of learned solutions is between 16.10-16.80, significantly worse than LKH3's 15.65). In this paper, we present ``Learn to Improve'' (L2I), the first learning based approach for CVRP that is efficient in solving speed and at the same time outperforms OR methods. Starting with a random initial solution, L2I learns to iteratively refine the solution with an improvement operator, selected by a reinforcement learning based controller. The improvement operator is selected from a pool of powerful operators that are customized for routing problems. By combining the strengths of the two worlds, our approach achieves the new state-of-the-art results on CVRP, e.g., an average cost of 15.57 on CVRP-100.
****

Transferable Perturbations of Deep Feature Distributions
Nathan Inkawhich,Kevin Liang,Lawrence Carin,Yiran Chen
Almost all current adversarial attacks of CNN classifiers rely on information derived from the output layer of the network. This work presents a new adversarial attack based on the modeling and exploitation of class-wise and layer-wise deep feature distributions. We achieve state-of-the-art targeted blackbox transfer-based attack results for undefended ImageNet models. Further, we place a priority on explainability and interpretability of the attacking process. Our methodology affords an analysis of how adversarial attacks change the intermediate feature distributions of CNNs, as well as a measure of layer-wise and class-wise feature distributional separability/entanglement. We also conceptualize a transition from task/data-specific to model-specific features within a CNN architecture that directly impacts the transferability of adversarial examples.
****

Skip Connections Matter: On the Transferability of Adversarial Examples Generated with ResNets
Dongxian Wu,Yisen Wang,Shu-Tao Xia,James Bailey,Xingjun Ma
Skip connections are an essential component of current state-of-the-art deep neural networks (DNNs) such as ResNet, WideResNet, DenseNet, and ResNeXt. Despite their huge success in building deeper and more powerful DNNs, we identify a surprising \emph{security weakness} of skip connections in this paper. Use of skip connections \textit{allows easier generation of highly transferable adversarial examples}. Specifically, in ResNet-like (with skip connections) neural networks, gradients can backpropagate through either skip connections or residual modules. We find that using more gradients from the skip connections rather than the residual modules according to a decay factor, allows one to craft adversarial examples with high transferability. Our method is termed \emph{Skip Gradient Method} (SGM). We conduct comprehensive transfer attacks against state-of-the-art DNNs including ResNets, DenseNets, Inceptions, Inception-ResNet, Squeeze-and-Excitation Network (SENet) and robustly trained DNNs. We show that employing SGM on the gradient flow can greatly improve the transferability of crafted attacks in almost all cases. Furthermore, SGM can be easily combined with existing black-box attack techniques, and obtain high improvements over state-of-the-art transferability methods. Our findings not only motivate new research into the architectural vulnerability of DNNs, but also open up further challenges for the design of secure DNN architectures.

****
ProtoAttend: Attention-Based Prototypical Learning
Sercan O. Arik,Tomas Pfister
We propose a novel inherently interpretable machine learning method that bases decisions on few relevant examples that we call prototypes. Our method, ProtoAttend, can be integrated into a wide range of neural network architectures including pre-trained models. It utilizes an attention mechanism that relates the encoded representations to samples in order to determine prototypes. The resulting model outperforms state of the art in three high impact problems without sacrificing accuracy of the original model: (1) it enables high-quality interpretability that outputs samples most relevant to the decision-making (i.e. a sample-based interpretability method); (2) it achieves state of the art confidence estimation by quantifying the mismatch across prototype labels; and (3) it obtains state of the art in distribution mismatch detection. All this can be achieved with minimal additional test time and a practically viable training time computational cost.
****
A Signal Propagation Perspective for Pruning Neural Networks at Initialization
Namhoon Lee,Thalaiyasingam Ajanthan,Stephen Gould,Philip H. S. Torr
Network pruning is a promising avenue for compressing deep neural networks. A typical approach to pruning starts by training a model and then removing redundant parameters while minimizing the impact on what is learned. Alternatively, a recent approach shows that pruning can be done at initialization prior to training, based on a saliency criterion called connection sensitivity. However, it remains unclear exactly why pruning an untrained, randomly initialized neural network is effective. In this work, by noting connection sensitivity as a form of gradient, we formally characterize initialization conditions to ensure reliable connection sensitivity measurements, which in turn yields effective pruning results. Moreover, we analyze the signal propagation properties of the resulting pruned networks and introduce a simple, data-free method to improve their trainability. Our modifications to the existing pruning at initialization method lead to improved results on all tested network models for image classification tasks. Furthermore, we empirically study the effect of supervision for pruning and demonstrate that our signal propagation perspective, combined with unsupervised pruning, can be useful in various scenarios where pruning is applied to non-standard arbitrarily-designed architectures.
****
Wildly Unsupervised Domain Adaptation and Its Powerful and Efficient Solution
Feng Liu,Jie Lu,Bo Han,Gang Niu,Guangquan Zhang,Masashi Sugiyama
In unsupervised domain adaptation (UDA), classifiers for the target domain (TD) are trained with clean labeled data from the source domain (SD) and unlabeled data from TD. However, in the wild, it is hard to acquire a large amount of perfectly clean labeled data in SD given limited budget. Hence, we consider a new, more realistic and more challenging problem setting, where classifiers have to be trained with noisy labeled data from SD and unlabeled data from TD---we name it wildly UDA (WUDA). We show that WUDA ruins all UDA methods if taking no care of label noise in SD, and to this end, we propose a Butterfly framework, a powerful and efficient solution to WUDA. Butterfly maintains four models (e.g., deep networks) simultaneously, where two take care of all adaptations (i.e., noisy-to-clean, labeled-to-unlabeled, and SD-to-TD-distributional) and then the other two can focus on classification in TD. As a consequence, Butterfly possesses all the conceptually necessary components for solving WUDA. Experiments demonstrate that under WUDA, Butterfly significantly outperforms existing baseline methods.
****
Automatically Learning Feature Crossing from Model Interpretation for Tabular Data
Zhaocheng Liu,Qiang Liu,Haoli Zhang
Automatically feature generation is a major topic of automated machine learning. Among various feature generation approaches, feature crossing, which takes cross-product of sparse features, is a promising way to effectively capture the inte

ractions among categorical features in tabular data. Previous works on feature c
rossing try to search in the set of all the possible cross feature fields. This
is obviously not efficient when the size of original feature fields is large. Me
anwhile, some deep learning-based methods combines deep neural networks and vari
ous interaction components. However, due to the existing of Deep Neural Networks
 (DNN), only a few cross features can be explicitly generated by the interaction
 components. Recently, piece-wise interpretation of DNN has been widely studied,
 and the piece-wise interpretations are usually inconsistent in different sample
s. Inspired by this, we give a definition of interpretation inconsistency in DNN
, and propose a novel method called CrossGO, which selects useful cross features
 according to the interpretation inconsistency. The whole process of learning fe
ature crossing can be done via simply training a DNN model and a logistic regres
sion (LR) model. CrossGO can generate compact candidate set of cross feature fie
lds, and promote the efficiency of searching. Extensive experiments have been co
nducted on several real-world datasets. Cross features generated by CrossGO can
empower a simple LR model achieving approximate or even better performances comp
aring with complex DNN models.
****

Continual Learning with Adaptive Weights (CLAW)
Tameem Adel,Han Zhao,Richard E. Turner
Approaches to continual learning aim to successfully learn a set of related task
s that arrive in an online manner. Recently, several frameworks have been develo
ped which enable deep learning to be deployed in this learning scenario. A key m
odelling decision is to what extent the architecture should be shared across tas
ks. On the one hand, separately modelling each task avoids catastrophic forgetti
ng but it does not support transfer learning and leads to large models. On the o
ther hand, rigidly specifying a shared component and a task-specific part enable
s task transfer and limits the model size, but it is vulnerable to catastrophic
 forgetting and restricts the form of task-transfer that can occur. Ideally, the
network should adaptively identify which parts of the network to share in a data
 driven way. Here we introduce such an approach called Continual Learning with A
daptive Weights (CLAW), which is based on probabilistic modelling and variationa
l inference. Experiments show that CLAW achieves state-of-the-art performance on
 six benchmarks in terms of overall continual learning performance, as measured
by classification accuracy, and in terms of addressing catastrophic forgetting.
****

Progressive Upsampling Audio Synthesis via Effective Adversarial Training
Youngwoo Cho,Minwook Chang,Gerard Jounghyun Kim,Jaegul Choo
This paper proposes a novel generative model called PUGAN, which progressively s
ynthesizes high-quality audio in a raw waveform. PUGAN leverages on the recently
 proposed idea of progressive generation of higher-resolution images by stacking
 multiple encode-decoder architectures. To effectively apply it to raw audio gen
eration, we propose two novel modules: (1) a neural upsampling layer and (2) a s
inc convolutional layer. Compared to the existing state-of-the-art model called
WaveGAN, which uses a single decoder architecture, our model generates audio sig
nals and converts them in a higher resolution in a progressive manner, while usi
ng a significantly smaller number of parameters, e.g., 20x smaller for 44.1kHz o
utput, than an existing technique called WaveGAN. Our experiments show that the
audio signals can be generated in real-time with the comparable quality to that
of WaveGAN with respect to the inception scores and the human evaluation.
****

Learning Compact Reward for Image Captioning
Nannan Li,Zhenzhong Chen
Adversarial learning has shown its advances in generating natural and diverse de
scriptions in image captioning. However, the learned reward of existing adversar
ial methods is vague and ill-defined due to the reward ambiguity problem. In thi
s paper, we propose a refined Adversarial Inverse Reinforcement Learning (rAIRL)
 method to handle the reward ambiguity problem by disentangling reward for each
word in a sentence, as well as achieve stable adversarial training by refining t
he loss function to shift the stationary point towards Nash equilibrium. In addi

tion, we introduce a conditional term in the loss function to mitigate mode collapse and to increase the diversity of the generated descriptions. Our experiments on MS COCO show that our method can learn compact reward for image captioning.
****

S-Flow GAN
Miron Yakov,Coscas Yona
Our work offers a new method for domain translation from semantic label maps and Computer Graphic (CG) simulation edge map images to photo-realistic images. We train a Generative Adversarial Network (GAN) in a conditional way to generate a photo-realistic version of a given CG scene. Existing architectures of GANs still lack the photo-realism capabilities needed to train DNNs for computer vision tasks, we address this issue by embedding edge maps, and training it in an adversarial mode. We also offer an extension to our model that uses our GAN architecture to create visually appealing and temporally coherent videos.
****

Gradient-free Neural Network Training by Multi-convex Alternating Optimization
Junxiang Wang,Fuxun Yu,Xiang Chen,Liang Zhao
In recent years, stochastic gradient descent (SGD) and its variants have been the dominant optimization methods for training deep neural networks. However, SGD suffers from limitations such as the lack of theoretical guarantees, vanishing gradients, excessive sensitivity to input, and difficulties solving highly non-smooth constraints and functions. To overcome these drawbacks, alternating minimization-based methods for deep neural network optimization have attracted fast-increasing attention recently. As an emerging and open domain, however, several new challenges need to be addressed, including 1) Convergence depending on the choice of hyperparameters, and 2) Lack of unified theoretical frameworks with general conditions. We, therefore, propose a novel Deep Learning Alternating Minimization (DLAM) algorithm to deal with these two challenges. Our innovative inequality-constrained formulation infinitely approximates the original problem with non-convex equality constraints, enabling our proof of global convergence of the DLAM algorithm under mild, practical conditions, regardless of the choice of hyperparameters and wide range of various activation functions. Experiments on benchmark datasets demonstrate the effectiveness of DLAM.
****

Semi-supervised Semantic Segmentation using Auxiliary Network
Wei-Hsu Chen,Hsueh-Ming Hang
Recently, the convolutional neural networks (CNNs) have shown great success on semantic segmentation task. However, for practical applications such as autonomous driving, the popular supervised learning method faces two challenges: the demand of low computational complexity and the need of huge training dataset accompanied by ground truth. Our focus in this paper is semi-supervised learning. We wish to use both labeled and unlabeled data in the training process. A highly efficient semantic segmentation network is our platform, which achieves high segmentation accuracy at low model size and high inference speed. We propose a semi-supervised learning approach to improve segmentation accuracy by including extra images without labels. While most existing semi-supervised learning methods are designed based on the adversarial learning techniques, we present a new and different approach, which trains an auxiliary CNN network that validates labels (ground-truth) on the unlabeled images. Therefore, in the supervised training phase, both the segmentation network and the auxiliary network are trained using labeled images. Then, in the unsupervised training phase, the unlabeled images are segmented and a subset of image pixels are picked up by the auxiliary network; and then they are used as ground truth to train the segmentation network. Thus, at the end, all dataset images can be used for retraining the segmentation network to improve the segmentation results. We use Cityscapes and CamVid datasets to verify the effectiveness of our semi-supervised scheme, and our experimental results show that it can improve the mean IoU for about 1.2% to 2.9% on the challenging Cityscapes dataset.

****
Intensity-Free Learning of Temporal Point Processes
Oleksandr Shchur,Marin Biloš,Stephan Günnemann
Temporal point processes are the dominant paradigm for modeling sequences of events happening at irregular intervals. The standard way of learning in such models is by estimating the conditional intensity function. However, parameterizing the intensity function usually incurs several trade-offs. We show how to overcome the limitations of intensity-based approaches by directly modeling the conditional distribution of inter-event times. We draw on the literature on normalizing flows to design models that are flexible and efficient. We additionally propose a simple mixture model that matches the flexibility of flow-based models, but also permits sampling and computing moments in closed form. The proposed models achieve state-of-the-art performance in standard prediction tasks and are suitable for novel applications, such as learning sequence embeddings and imputing missing data.
****
Scalable and Order-robust Continual Learning with Additive Parameter Decomposition
Jaehong Yoon,Saehoon Kim,Eunho Yang,Sung Ju Hwang
While recent continual learning methods largely alleviate the catastrophic problem on toy-sized datasets, there are issues that remain to be tackled in order to apply them to real-world problem domains. First, a continual learning model should effectively handle catastrophic forgetting and be efficient to train even with a large number of tasks. Secondly, it needs to tackle the problem of order-sensitivity, where the performance of the tasks largely varies based on the order of the task arrival sequence, as it may cause serious problems where fairness plays a critical role (e.g. medical diagnosis). To tackle these practical challenges, we propose a novel continual learning method that is scalable as well as order-robust, which instead of learning a completely shared set of weights, represents the parameters for each task as a sum of task-shared and sparse task-adaptive parameters. With our Additive Parameter Decomposition (APD), the task-adaptive parameters for earlier tasks remain mostly unaffected, where we update them only to reflect the changes made to the task-shared parameters. This decomposition of parameters effectively prevents catastrophic forgetting and order-sensitivity, while being computation- and memory-efficient. Further, we can achieve even better scalability with APD using hierarchical knowledge consolidation, which clusters the task-adaptive parameters to obtain hierarchically shared parameters. We validate our network with APD, APD-Net, on multiple benchmark datasets against state-of-the-art continual learning methods, which it largely outperforms in accuracy, scalability, and order-robustness.
****
Discriminator Based Corpus Generation for General Code Synthesis
Alexander Wild,Barry Porter
Current work on neural code synthesis consists of increasingly sophisticated architectures being trained on highly simplified domain-specific languages, using uniform sampling across program space of those languages for training. By comparison, program space for a C-like language is vast, and extremely sparsely populated in terms of `useful' functionalities; this requires a far more intelligent approach to corpus generation for effective training. We use a genetic programming approach using an iteratively retrained discriminator to produce a population suitable as labelled training data for a neural code synthesis architecture. We demonstrate that use of a discriminator-based training corpus generator, trained using only unlabelled problem specifications in classic Programming-by-Example format, greatly improves network performance compared to current uniform sampling techniques.
****
Storage Efficient and Dynamic Flexible Runtime Channel Pruning via Deep Reinforcement Learning
Jianda Chen,Shangyu Chen,Sinno Jialin Pan
In this paper, we propose a deep reinforcement learning (DRL) based framework to

efficiently perform runtime channel pruning on convolutional neural networks (CNNs). Our DRL-based framework aims to learn a pruning strategy to determine how many and which channels to be pruned in each convolutional layer, depending on each specific input instance in runtime. The learned policy optimizes the performance of the network by restricting the computational resource on layers under an overall computation budget. Furthermore, unlike other runtime pruning methods which require to store all channels parameters in inference, our framework can reduce parameters storage consumption at deployment by introducing a static pruning component. Comparison experimental results with existing runtime and static pruning methods on state-of-the-art CNNs demonstrate that our proposed framework is able to provide a tradeoff between dynamic flexibility and storage efficiency in runtime channel pruning.

****

BOOSTING ENCODER-DECODER CNN FOR INVERSE PROBLEMS

Eunju Cha,Jaeduck Jang,Junho Lee,Eunha Lee,Jong Chul Ye

Encoder-decoder convolutional neural networks (CNN) have been extensively used for various inverse problems. However, their prediction error for unseen test data is difficult to estimate a priori, since the neural networks are trained using only selected data and their architectures are largely considered blackboxes. This poses a fundamental challenge in improving the performance of neural networks. Recently, it was shown that Stein's unbiased risk estimator (SURE) can be used as an unbiased estimator of the prediction error for denoising problems. However, the computation of the divergence term in SURE is difficult to implement in a neural network framework, and the condition to avoid trivial identity mapping is not well defined. In this paper, inspired by the finding that an encoder-decoder CNN can be expressed as a piecewise linear representation, we provide a close form expression of the unbiased estimator for the prediction error. The close form representation leads to a novel boosting scheme to prevent a neural network from converging to an identity mapping so that it can enhance the performance. Experimental results show that the proposed algorithm provides consistent improvement in various inverse problems.

****

Weakly Supervised Clustering by Exploiting Unique Class Count

Mustafa Umit Oner,Hwee Kuan Lee,Wing-Kin Sung

A weakly supervised learning based clustering framework is proposed in this paper. As the core of this framework, we introduce a novel multiple instance learning task based on a bag level label called unique class count (ucc), which is the number of unique classes among all instances inside the bag. In this task, no annotations on individual instances inside the bag are needed during training of the models. We mathematically prove that with a perfect ucc classifier, perfect clustering of individual instances inside the bags is possible even when no annotations on individual instances are given during training. We have constructed a neural network based ucc classifier and experimentally shown that the clustering performance of our framework with our weakly supervised ucc classifier is comparable to that of fully supervised learning models where labels for all instances are known. Furthermore, we have tested the applicability of our framework to a real world task of semantic segmentation of breast cancer metastases in histological lymph node sections and shown that the performance of our weakly supervised framework is comparable to the performance of a fully supervised Unet model.

****

Domain Adaptation via Low-Rank Basis Approximation

Christoph Raab,Frank-Michael Schleif

Domain adaptation focuses on the reuse of supervised learning models in a new context. Prominent applications can be found in robotics, image processing or web mining. In these areas, learning scenarios change by nature, but often remain related and motivate the reuse of existing supervised models.

While the majority of symmetric and asymmetric domain adaptation algorithms utilize all available source and target domain data, we show that efficient domain adaptation requires only a substantially smaller subset from both domains. This makes it more suitable for real-world scenarios where target domain data is rare.

The presented approach finds a target subspace representation for source and ta
rget data to address domain differences by orthogonal basis transfer. By employi
ng a low-rank approximation, the approach remains low in computational time.
The presented idea is evaluated in typical domain adaptation tasks with standard
 benchmark data.
****
Learning to Control PDEs with Differentiable Physics
Philipp Holl,Nils Thuerey,Vladlen Koltun
Predicting outcomes and planning interactions with the physical world are long-s
tanding goals for machine learning. A variety of such tasks involves continuous
physical systems, which can be described by partial differential equations (PDEs
) with many degrees of freedom. Existing methods that aim to control the dynamic
s of such systems are typically limited to relatively short time frames or a sma
ll number of interaction parameters. We present a novel hierarchical predictor-c
orrector scheme which enables neural networks to learn to understand and control
 complex nonlinear physical systems over long time frames. We propose to split t
he problem into two distinct tasks: planning and control. To this end, we introd
uce a predictor network that plans optimal trajectories and a control network th
at infers the corresponding control parameters. Both stages are trained end-to-e
nd using a differentiable PDE solver. We demonstrate that our method successfull
y develops an understanding of complex physical systems and learns to control th
em for tasks involving PDEs such as the incompressible Navier-Stokes equations.
****
Linear Symmetric Quantization of Neural Networks for Low-precision Integer Hardw
are
Xiandong Zhao,Ying Wang,Xuyi Cai,Cheng Liu,Lei Zhang
With the proliferation of specialized neural network processors that operate on
low-precision integers, the performance of Deep Neural Network inference becomes
 increasingly dependent on the result of quantization. Despite plenty of prior w
ork on the quantization of weights or activations for neural networks, there is
still a wide gap between the software quantizers and the low-precision accelerat
or implementation, which degrades either the efficiency of networks or that of t
he hardware for the lack of software and hardware coordination at design-phase.
In this paper, we propose a learned linear symmetric quantizer for integer neura
l network processors, which not only quantizes neural parameters and activations
 to low-bit integer but also accelerates hardware inference by using batch norma
lization fusion and low-precision accumulators (e.g., 16-bit) and multipliers (e
.g., 4-bit). We use a unified way to quantize weights and activations, and the r
esults outperform many previous approaches for various networks such as AlexNet,
 ResNet, and lightweight models like MobileNet while keeping friendly to the acc
elerator architecture. Additional, we also apply the method to object detection
models and witness high performance and accuracy in YOLO-v2. Finally, we deploy
the quantized models on our specialized integer-arithmetic-only DNN accelerator
to show the effectiveness of the proposed quantizer. We show that even with line
ar symmetric quantization, the results can be better than asymmetric or non-line
ar methods in 4-bit networks. In evaluation, the proposed quantizer induces less
 than 0.4\% accuracy drop in ResNet18, ResNet34, and AlexNet when quantizing the
 whole network as required by the integer processors.
****
Estimating Gradients for Discrete Random Variables by Sampling without Replaceme
nt
Wouter Kool,Herke van Hoof,Max Welling
We derive an unbiased estimator for expectations over discrete random variables
based on sampling without replacement, which reduces variance as it avoids dupli
cate samples. We show that our estimator can be derived as the Rao-Blackwellizat
ion of three different estimators. Combining our estimator with REINFORCE, we ob
tain a policy gradient estimator and we reduce its variance using a built-in con
trol variate which is obtained without additional model evaluations. The resulti
ng estimator is closely related to other gradient estimators. Experiments with a
 toy problem, a categorical Variational Auto-Encoder and a structured prediction

problem show that our estimator is the only estimator that is consistently among the best estimators in both high and low entropy settings.
****
On importance-weighted autoencoders
Axel Finke,Alexandre H. Thiery
The importance weighted autoencoder (IWAE) (Burda et al., 2016) is a popular variational-inference method which achieves a tighter evidence bound (and hence a lower bias) than standard variational autoencoders by optimising a multi-sample objective, i.e. an objective that is expressible as an integral over $K > 1$ Monte Carlo samples. Unfortunately, IWAE crucially relies on the availability of reparametrisations and even if these exist, the multi-sample objective leads to inference-network gradients which break down as K is increased (Rainforth et al., 2018). This breakdown can only be circumvented by removing high-variance score-function terms, either by heuristically ignoring them (which yields the 'sticking-the-landing' IWAE (IWAE-STL) gradient from Roeder et al. (2017)) or through an identity from Tucker et al. (2019) (which yields the 'doubly-reparametrised' IWAE (IWAE-DREG) gradient). In this work, we argue that directly optimising the proposal distribution in importance sampling as in the reweighted wake-sleep (RWS) algorithm from Bornschein & Bengio (2015) is preferable to optimising IWAE-type multi-sample objectives. To formalise this argument, we introduce an adaptive-importance sampling framework termed adaptive importance sampling for learning (AISLE) which slightly generalises the RWS algorithm. We then show that AISLE admits IWAE-STL and IWAE-DREG (i.e. the IWAE-gradients which avoid breakdown) as special cases.
****
FALCON: Fast and Lightweight Convolution for Compressing and Accelerating CNN
Chun Quan,Jun-Gi Jang,Hyun Dong Lee,U Kang
How can we efficiently compress Convolutional Neural Networks (CNN) while retaining their accuracy on classification tasks? A promising direction is based on depthwise separable convolution which replaces a standard convolution with a depthwise convolution and a pointwise convolution. However, previous works based on depthwise separable convolution are limited since 1) they are mostly heuristic approaches without a precise understanding of their relations to standard convolution, and 2) their accuracies do not match that of the standard convolution.

In this paper, we propose FALCON, an accurate and lightweight method for compressing CNN. FALCON is derived by interpreting existing convolution methods based on depthwise separable convolution using EHP, our proposed mathematical formulation to approximate the standard convolution kernel. Such interpretation leads to developing a generalized version rank-k FALCON which further improves the accuracy while sacrificing a bit of compression and computation reduction rates. In addition, we propose FALCON-branch by fitting FALCON into the previous state-of-the-art convolution unit ShuffleUnitV2 which gives even better accuracy. Experiments show that FALCON and FALCON-branch outperform 1) existing methods based on depthwise separable convolution and 2) standard CNN models by up to 8x compression and 8x computation reduction while ensuring similar accuracy. We also demonstrate that rank-k FALCON provides even better accuracy than standard convolution in many cases, while using a smaller number of parameters and floating-point operations.
****
Mincut Pooling in Graph Neural Networks
Filippo Maria Bianchi,Daniele Grattarola,Cesare Alippi
The advance of node pooling operations in Graph Neural Networks (GNNs) has lagged behind the feverish design of new message-passing techniques, and pooling remains an important and challenging endeavor for the design of deep architectures.
In this paper, we propose a pooling operation for GNNs that leverages a differentiable unsupervised loss based on the minCut optimization objective.
For each node, our method learns a soft cluster assignment vector that depends on the node features, the target inference task (e.g., a graph classification loss), and, thanks to the minCut objective, also on the connectivity structure of t

he graph.
Graph pooling is obtained by applying the matrix of assignment vectors to the ad jacency matrix and the node features.
We validate the effectiveness of the proposed pooling method on a variety of sup ervised and unsupervised tasks.
****
Dual Graph Representation Learning
Huiling Zhu,Xin Luo,Hankz Hankui Zhuo
Graph representation learning embeds nodes in large graphs as low-dimensional ve ctors and benefit to many downstream applications. Most embedding frameworks, h owever, are inherently transductive and unable to generalize to unseen nodes or learn representations across different graphs. Inductive approaches, such as Gr aphSAGE, neglect different contexts of nodes and cannot learn node embeddings d ually. In this paper, we present an unsupervised dual encoding framework, \textb f{CADE}, to generate context-aware representation of nodes by combining real-ti me neighborhood structure with neighbor-attentioned representation, and preservi ng extra memory of known nodes. Experimently, we exhibit that our approach is ef fective by comparing to state-of-the-art methods.
****
Unsupervised Few Shot Learning via Self-supervised Training
Zilong Ji,Xiaolong Zou,Tiejun Huang,Si Wu
Learning from limited exemplars (few-shot learning) is a fundamental, unsolved p roblem that has been laboriously explored in the machine learning community. How ever, current few-shot learners are mostly supervised and rely heavily on a larg e amount of labeled examples. Unsupervised learning is a more natural procedure for cognitive mammals and has produced promising results in many machine learnin g tasks. In the current study, we develop a method to learn an unsupervised few- shot learner via self-supervised training (UFLST), which can effectively general ize to novel but related classes. The proposed model consists of two alternate p rocesses, progressive clustering and episodic training. The former generates pse udo-labeled training examples for constructing episodic tasks; and the later tr ains the few-shot learner using the generated episodic tasks which further optim izes the feature representations of data. The two processes facilitate with each other, and eventually produce a high quality few-shot learner. Using the benchm ark dataset Omniglot, we show that our model outperforms other unsupervised few- shot learning methods to a large extend and approaches to the performances of su pervised methods. Using the benchmark dataset Market1501, we further demonstrate the feasibility of our model to a real-world application on person re-identific ation.
****
To Relieve Your Headache of Training an MRF, Take AdVIL
Chongxuan Li,Chao Du,Kun Xu,Max Welling,Jun Zhu,Bo Zhang
We propose a black-box algorithm called {\it Adversarial Variational Inference a nd Learning} (AdVIL) to perform inference and learning on a general Markov rand om field (MRF). AdVIL employs two variational distributions to approximately inf er the latent variables and estimate the partition function of an MRF, respectiv ely. The two variational distributions provide an estimate of the negative log-l ikelihood of the MRF as a minimax optimization problem, which is solved by stoch astic gradient descent. AdVIL is proven convergent under certain conditions. On one hand, compared with contrastive divergence, AdVIL requires a minimal assumpt ion about the model structure and can deal with a broader family of MRFs. On the other hand, compared with existing black-box methods, AdVIL provides a tighter estimate of the log partition function and achieves much better empirical result s.
****
On the Dynamics and Convergence of Weight Normalization for Training Neural Netw orks
Yonatan Dukler,Quanquan Gu,Guido Montufar
We present a proof of convergence for ReLU networks trained with weight normaliz ation. In the analysis, we consider over-parameterized 2-layer ReLU networks ini

tialized at random and trained with batch gradient descent and a fixed step size . The proof builds on recent theoretical works that bound the trajectory of parameters from their initialization and monitor the network predictions via the evolution of a ''neural tangent kernel'' (Jacot et al. 2018). We discover that training with weight normalization decomposes such a kernel via the so called ''length-direction decoupling''. This in turn leads to two convergence regimes and can rigorously explain the utility of WeightNorm. From the modified convergence we make a few curious observations including a natural form of ''lazy training'' where the direction of each weight vector remains stationary.
****

Understanding and Improving Transformer From a Multi-Particle Dynamic System Point of View
Yiping Lu,Zhuohan Li,Di He,Zhiqing Sun,Bin Dong,Tao Qin,Liwei Wang,Tie-Yan Liu
The Transformer architecture is widely used in natural language processing. Despite its success, the design principle of the Transformer remains elusive. In this paper, we provide a novel perspective towards understanding the architecture: we show that the Transformer can be mathematically interpreted as a numerical Ordinary Differential Equation (ODE) solver for a convection-diffusion equation in a multi-particle dynamic system. In particular, how words in a sentence are abstracted into contexts by passing through the layers of the Transformer can be interpreted as approximating multiple particles' movement in the space using the Lie-Trotter splitting scheme and the Euler's method. Given this ODE's perspective, the rich literature of numerical analysis can be brought to guide us in designing effective structures beyond the Transformer. As an example, we propose to replace the Lie-Trotter splitting scheme by the Strang-Marchuk splitting scheme, a scheme that is more commonly used and with much lower local truncation errors. The Strang-Marchuk splitting scheme suggests that the self-attention and position-wise feed-forward network (FFN) sub-layers should not be treated equally. Instead, in each layer, two position-wise FFN sub-layers should be used, and the self-attention sub-layer is placed in between. This leads to a brand new architecture. Such an FFN-attention-FFN layer is "Macaron-like", and thus we call the network with this new architecture the Macaron Net. Through extensive experiments, we show that the Macaron Net is superior to the Transformer on both supervised and unsupervised learning tasks. The reproducible code can be found on http://anonymized
****

SesameBERT: Attention for Anywhere
Ta-Chun Su,Hsiang-Chih Cheng
Fine-tuning with pre-trained models has achieved exceptional results for many language tasks. In this study, we focused on one such self-attention network model, namely BERT, which has performed well in terms of stacking layers across diverse language-understanding benchmarks. However, in many downstream tasks, information between layers is ignored by BERT for fine-tuning. In addition, although self-attention networks are well-known for their ability to capture global dependencies, room for improvement remains in terms of emphasizing the importance of local contexts. In light of these advantages and disadvantages, this paper proposes SesameBERT, a generalized fine-tuning method that (1) enables the extraction of global information among all layers through Squeeze and Excitation and (2) enriches local information by capturing neighboring contexts via Gaussian blurring. Furthermore, we demonstrated the effectiveness of our approach in the HANS dataset, which is used to determine whether models have adopted shallow heuristics instead of learning underlying generalizations. The experiments revealed that SesameBERT outperformed BERT with respect to GLUE benchmark and the HANS evaluation set.
****

Automated Relational Meta-learning
Huaxiu Yao,Xian Wu,Zhiqiang Tao,Yaliang Li,Bolin Ding,Ruirui Li,Zhenhui Li
In order to efficiently learn with small amount of data on new tasks, meta-learning transfers knowledge learned from previous tasks to the new ones. However, a critical challenge in meta-learning is the task heterogeneity which cannot be we

ll handled by traditional globally shared meta-learning methods. In addition, cu rrent task-specific meta-learning methods may either suffer from hand-crafted st ructure design or lack the capability to capture complex relations between tasks . In this paper, motivated by the way of knowledge organization in knowledge bas es, we propose an automated relational meta-learning (ARML) framework that autom atically extracts the cross-task relations and constructs the meta-knowledge gra ph. When a new task arrives, it can quickly find the most relevant structure and tailor the learned structure knowledge to the meta-learner. As a result, the pr oposed framework not only addresses the challenge of task heterogeneity by a lea rned meta-knowledge graph, but also increases the model interpretability. We con duct extensive experiments on 2D toy regression and few-shot image classificatio n and the results demonstrate the superiority of ARML over state-of-the-art base lines.

****

Training Deep Networks with Stochastic Gradient Normalized by Layerwise Adaptive Second Moments

Boris Ginsburg,Patrice Castonguay,Oleksii Hrinchuk,Oleksii Kuchaiev,Vitaly Lavru khin,Ryan Leary,Jason Li,Huyen Nguyen,Yang Zhang,Jonathan M. Cohen

We propose NovoGrad, an adaptive stochastic gradient descent method with layer-w ise gradient normalization and decoupled weight decay. In our experiments on neu ral networks for image classification, speech recognition, machine translation, and language modeling, it performs on par or better than well tuned SGD with mom entum and Adam/AdamW.

Additionally, NovoGrad (1) is robust to the choice of learning rate and weight i nitialization, (2) works well in a large batch setting, and (3) has two times sm aller memory footprint than Adam.

****

Moniqua: Modulo Quantized Communication in Decentralized SGD

Yucheng Lu,Christopher De Sa

Decentralized stochastic gradient descent (SGD), where parallel workers are conn ected to form a graph and communicate adjacently, has shown promising results bo th theoretically and empirically. In this paper we propose Moniqua, a technique that allows decentralized SGD to use quantized communication. We prove in theory that Moniqua communicates a provably bounded number of bits per iteration, whil e converging at the same asymptotic rate as the original algorithm does with ful l-precision communication. Moniqua improves upon prior works in that it (1) requ ires no additional memory, (2) applies to non-convex objectives, and (3) support s biased/linear quantizers. We demonstrate empirically that Moniqua converges fa ster with respect to wall clock time than other quantized decentralized algorith ms. We also show that Moniqua is robust to very low bit-budgets, allowing less than 4-bits-per-parameter communication without affecting convergence when trai ning VGG16 on CIFAR10.

****

Defending Against Physically Realizable Attacks on Image Classification

Tong Wu,Liang Tong,Yevgeniy Vorobeychik

We study the problem of defending deep neural network approaches for image class ification from physically realizable attacks. First, we demonstrate that the two most scalable and effective methods for learning robust models, adversarial tra ining with PGD attacks and randomized smoothing, exhibit very limited effectiven ess against three of the highest profile physical attacks. Next, we propose a ne w abstract adversarial model, rectangular occlusion attacks, in which an adversa ry places a small adversarially crafted rectangle in an image, and develop two a pproaches for efficiently computing the resulting adversarial examples. Finally, we demonstrate that adversarial training using our new attack yields image clas sification models that exhibit high robustness against the physically realizable attacks we study, offering the first effective generic defense against such att acks.

****

Certifying Distributional Robustness using Lipschitz Regularisation

Zac Cranko,Zhan Shi,Xinhua Zhang,Simon Kornblith,Richard Nock

Distributional robust risk (DRR) minimisation has arisen as a flexible and effec
tive framework for machine learning. Approximate solutions based on dualisation
have become particularly favorable in addressing the semi-infinite optimisation,
and they also provide a certificate of the robustness for the worst-case popula
tion loss. However existing methods are restricted to either linear models or ve
ry small perturbations, and cannot find the globally optimal solution for restri
cted nonlinear models such as kernel methods. In this paper we resolved these li
mitations by upper bounding DRRs with an empirical risk regularised by the Lipsc
hitz constant of the model, including deep neural networks and kernel methods. A
s an application, we showed that it also provides a certificate for adversarial
training, and global solutions can be achieved on product kernel machines in pol
ynomial time.
****

A SPIKING SEQUENTIAL MODEL: RECURRENT LEAKY INTEGRATE-AND-FIRE
Daiheng Gao,Hongwei Wang,Hehui Zhang,Meng Wang,Zhenzhi Wu
Stemming from neuroscience, Spiking neural networks (SNNs), a brain-inspired neu
ral network that is a versatile solution to fault-tolerant and energy efficient
information processing pertains to the ”event-driven” characteristic as the anal
ogy of the behavior of biological neurons. However, they are inferior to artific
ial neural networks (ANNs) in real complicated tasks and only had it been achiev
ed good results in rather simple applications. When ANNs usually being questione
d about it expensive processing costs and lack of essential biological plausibil
ity, the temporal characteristic of RNN-based architecture makes it suitable to
incorporate SNN inside as imitating the transition of membrane potential through
time, and a brain-inspired Recurrent Leaky Integrate-and-Fire (RLIF) model has
been put forward to overcome a series of challenges, such as discrete binary out
put and dynamical trait. The experiment results show that our recurrent architec
ture has an ultra anti-interference ability and strictly follows the guideline o
f SNN that spike output through it is discrete. Furthermore, this architecture a
chieves a good result on neuromorphic datasets and can be extended to tasks like
text summarization and video understanding.
****

N-BEATS: Neural basis expansion analysis for interpretable time series forecasti
ng
Boris N. Oreshkin,Dmitri Carpov,Nicolas Chapados,Yoshua Bengio
We focus on solving the univariate times series point forecasting problem using
deep learning. We propose a deep neural architecture based on backward and forwa
rd residual links and a very deep stack of fully-connected layers. The architect
ure has a number of desirable properties, being interpretable, applicable withou
t modification to a wide array of target domains, and fast to train. We test the
proposed architecture on several well-known datasets, including M3, M4 and TOUR
ISM competition datasets containing time series from diverse domains. We demonst
rate state-of-the-art performance for two configurations of N-BEATS for all the
datasets, improving forecast accuracy by 11% over a statistical benchmark and by
3% over last year's winner of the M4 competition, a domain-adjusted hand-crafte
d hybrid between neural network and statistical time series models. The first co
nfiguration of our model does not employ any time-series-specific components and
its performance on heterogeneous datasets strongly suggests that, contrarily to
received wisdom, deep learning primitives such as residual blocks are by themse
lves sufficient to solve a wide range of forecasting problems. Finally, we demon
strate how the proposed architecture can be augmented to provide outputs that ar
e interpretable without considerable loss in accuracy.
****

Subgraph Attention for Node Classification and Hierarchical Graph Pooling
Sambaran Bandyopadhyay,Manasvi Aggarwal,M. N. Murty
Graph neural networks have gained significant interest from the research communi
ty for both node classification within a graph and graph classification within a
set of graphs. Attention mechanism applied on the neighborhood of a node improv
es the performance of graph neural networks. Typically, it helps to identify a n
eighbor node which plays more important role to determine the label of the node

under consideration. But in real world scenarios, a particular subset of nodes t
ogether, but not the individual nodes in the subset, may be important to determi
ne the label of a node. To address this problem, we introduce the concept of sub
graph attention for graphs. To show the efficiency of this, we use subgraph atte
ntion with graph convolution for node classification. We further use subgraph at
tention for the entire graph classification by proposing a novel hierarchical ne
ural graph pooling architecture. Along with attention over the subgraphs, our po
oling architecture also uses attention to determine the important nodes within a
 level graph and attention to determine the important levels in the whole hierar
chy. Competitive performance over the state-of-the-arts for both node and graph
classification shows the efficiency of the algorithms proposed in this paper.
****

Are there any 'object detectors' in the hidden layers of CNNs trained to identif
y objects or scenes?

Ella M. Gale,Nicholas Martin,Ryan Blything,Anh Nguyen,Jeffrey S. Bowers

Various methods of measuring unit selectivity have been developed with the aim o
f better understanding how neural networks work. But the different measures pro
vide divergent estimates of selectivity, and this has led to different conclusio
ns regarding the conditions in which selective object representations are learne
d and the functional relevance of these representations. In an attempt to better
 characterize object selectivity, we undertake a comparison of various selectivi
ty measures on a large set of units in AlexNet, including localist selectivity,
precision, class-conditional mean activity selectivity (CCMAS), network dissecti
on, the human interpretation of activation maximization (AM) images, and standar
d signal-detection measures. We find that the different measures provide differ
ent estimates of object selectivity, with precision and CCMAS measures providing
 misleadingly high estimates. Indeed, the most selective units had a poor hit-ra
te or a high false-alarm rate (or both) in object classification, making them po
or object detectors. We fail to find any units that are even remotely as select
ive as the 'grandmother cell' units reported in recurrent neural networks. In or
der to generalize these results, we compared selectivity measures on a few units
 in VGG-16 and GoogLeNet trained on the ImageNet or Places-365 datasets that hav
e been described as 'object detectors'. Again, we find poor hit-rates and high f
alse-alarm rates for object classification.
****

Learning Human Postural Control with Hierarchical Acquisition Functions

Nils Rottmann,Tjasa Kunavar,Jan Babic,Jan Peters,Elmar Rueckert

Learning control policies in robotic tasks requires a large number of interactio
ns due to small learning rates, bounds on the updates or unknown constraints. In
 contrast humans can infer protective and safe solutions after a single failure
or unexpected observation.

In order to reach similar performance, we developed a hierarchical Bayesian opti
mization algorithm that replicates the cognitive inference and memorization proc
ess for avoiding failures in motor control tasks. A Gaussian Process implements
the modeling and the sampling of the acquisition function. This enables rapid le
arning with large learning rates while a mental replay phase ensures that policy
 regions that led to failures are inhibited during the sampling process.

The features of the hierarchical Bayesian optimization method are evaluated in a
 simulated and physiological humanoid postural balancing task. We quantitatively
 compare the human learning performance to our learning approach by evaluating t
he deviations of the center of mass during training. Our results show that we ca
n reproduce the efficient learning of human subjects in postural control tasks w
hich provides a testable model for future physiological motor control tasks. In
these postural control tasks, our method outperforms standard Bayesian Optimizat
ion in the number of interactions to solve the task, in the computational demand
s and in the frequency of observed failures.
****

Unsupervised Intuitive Physics from Past Experiences

Sebastien Ehrhardt,Aron Monszpart,Niloy Mitra,Andrea Vedaldi

We consider the problem of learning models of intuitive physics from raw, unlabe

lled visual input. Differently from prior work, in addition to learning general physical principles, we are also interested in learning ``on the fly'' physical properties specific to new environments, based on a small number of environment-specific experiences. We do all this in an unsupervised manner, using a meta-lea rning formulation where the goal is to predict videos containing demonstrations of physical phenomena, such as objects moving and colliding with a complex backg round. We introduce the idea of summarizing past experiences in a very compact m anner, in our case using dynamic images, and show that this can be used to solve the problem well and efficiently. Empirically, we show, via extensive experimen ts and ablation studies, that our model learns to perform physical predictions t hat generalize well in time and space, as well as to a variable number of intera cting physical objects.
****

Expected Tight Bounds for Robust Deep Neural Network Training
Salman Alsubaihi,Adel Bibi,Modar Alfadly,Abdullah Hamdi,Bernard Ghanem
Training Deep Neural Networks (DNNs) that are robust to norm bounded adversarial attacks remains an elusive problem. While verification based methods are genera lly too expensive to robustly train large networks, it was demonstrated by Gowal et. al. that bounded input intervals can be inexpensively propagated from layer to layer through deep networks. This interval bound propagation (IBP) approach led to high robustness and was the first to be employed on large networks. Howev er, due to the very loose nature of the IBP bounds, particularly for large/deep networks, the required training procedure is complex and involved. In this paper , we closely examine the bounds of a block of layers composed of an affine layer , followed by a ReLU, followed by another affine layer. To this end, we propose \emph{expected} bounds (true bounds in expectation), which are provably tighter than IBP bounds in expectation. We then extend this result to deeper networks th rough blockwise propagation and show that we can achieve orders of magnitudes ti ghter bounds compared to IBP. Using these tight bounds, we demonstrate that a si mple standard training procedure can achieve impressive robustness-accuracy trad e-off across several architectures on both MNIST and CIFAR10.
****

Analytical Moment Regularizer for Training Robust Networks
Modar Alfadly,Adel Bibi,Muhammed Kocabas,Bernard Ghanem
Despite the impressive performance of deep neural networks (DNNs) on numerous le arning tasks, they still exhibit uncouth behaviours. One puzzling behaviour is the subtle sensitive reaction of DNNs to various noise attacks. Such a nuisance has strengthened the line of research around developing and training noise-robus t networks. In this work, we propose a new training regularizer that aims to min imize the probabilistic expected training loss of a DNN subject to a generic Gau ssian input. We provide an efficient and simple approach to approximate such a r egularizer for arbitrarily deep networks. This is done by leveraging the analyti c expression of the output mean of a shallow neural network, avoiding the need f or memory and computation expensive data augmentation. We conduct extensive expe riments on LeNet and AlexNet on various datasets including MNIST, CIFAR10, and C IFAR100 to demonstrate the effectiveness of our proposed regularizer. In particu lar, we show that networks that are trained with the proposed regularizer benefi t from a boost in robustness against Gaussian noise to an equivalent amount of p erforming 3-21 folds of noisy data augmentation. Moreover, we empirically show o n several architectures and datasets that improving robustness against Gaussian noise, by using the new regularizer, can improve the overall robustness against 6 other types of attacks by two orders of magnitude.
****

Model Architecture Controls Gradient Descent Dynamics: A Combinatorial Path-Base d Formula
Xin Zhou,Newsha Ardalani
Recently, there has been a growing interest in automatically exploring neural ne twork architecture design space with the goal of finding an architecture that im proves performance (characterized as improved accuracy, speed of training, or re source requirements). However, our theoretical understanding of how model archit

ecture affects performance or accuracy is limited. In this paper, we study the i
mpact of model architecture on the speed of training in the context of gradient
descent optimization. We model gradient descent as a first-order ODE and use ODE
's coefficient matrix H to characterize the convergence rate. We introduce a sim
ple analysis technique that enumerates H in terms of all possible ``paths'' in t
he network.
 We show that changes in model architecture parameters reflect as changes in the
 number of paths and the properties of each path, which jointly control the spee
d of convergence. We believe our analysis technique is useful in reasoning about
 more complex model architecture modifications.
****

Deep Learning of Determinantal Point Processes via Proper Spectral Sub-gradient
Tianshu Yu,Yikang Li,Baoxin Li
Determinantal point processes (DPPs) is an effective tool to deliver diversity o
n multiple machine learning and computer vision tasks. Under deep learning frame
work, DPP is typically optimized via approximation, which is not straightforward
 and has some conflict with diversity requirement. We note, however, there has b
een no deep learning paradigms to optimize DPP directly since it involves matrix
 inversion which may result in highly computational instability. This fact great
ly hinders the wide use of DPP on some specific objectives where DPP serves as a
 term to measure the feature diversity. In this paper, we devise a simple but ef
fective algorithm to address this issue to optimize DPP term directly expressed
with L-ensemble in spectral domain over gram matrix, which is more flexible than
 learning on parametric kernels. By further taking into account some geometric c
onstraints, our algorithm seeks to generate valid sub-gradients of DPP term in c
ase when the DPP gram matrix is not invertible (no gradients exist in this case)
. In this sense, our algorithm can be easily incorporated with multiple deep lea
rning tasks. Experiments show the effectiveness of our algorithm, indicating pro
mising performance for practical learning problems.
****

Collaborative Filtering With A Synthetic Feedback Loop
Wenlin Wang,Hongteng Xu,Ruiyi Zhang,Wenqi Wang,Lawrence Carin
We propose a novel learning framework for recommendation systems, assisting coll
aborative filtering with a synthetic feedback loop. The proposed framework consi
sts of a ``recommender'' and a ``virtual user.'' The recommender is formulizd as
 a collaborative-filtering method, recommending items according to observed user
 behavior. The virtual user estimates rewards from the recommended items and ge
nerates the influence of the rewards on observed user behavior. The recommender
 connected with the virtual user constructs a closed loop, that recommends users
 with items and imitates the unobserved feedback of the users to the recommended
 items. The synthetic feedback is used to augment observed user behavior and imp
rove recommendation results. Such a model can be interpreted as the inverse rei
nforcement learning, which can be learned effectively via rollout (simulation).
Experimental results show that the proposed framework is able to boost the perfo
rmance of existing collaborative filtering methods on multiple datasets.
****

Self-Supervised State-Control through Intrinsic Mutual Information Rewards
Rui Zhao,Volker Tresp,Wei Xu
Learning to discover useful skills without a manually-designed reward function w
ould have many applications, yet is still a challenge for reinforcement learning
. In this paper, we propose Mutual Information-based State-Control (MISC), a new
 self-supervised Reinforcement Learning approach for learning to control states
of interest without any external reward function. We formulate the intrinsic obj
ective as rewarding the skills that maximize the mutual information between the
context states and the states of interest. For example, in robotic manipulation
tasks, the context states are the robot states and the states of interest are th
e states of an object. We evaluate our approach for different simulated robotic
manipulation tasks from OpenAI Gym. We show that our method is able to learn to
manipulate the object, such as pushing and picking up, purely based on the intri
nsic mutual information rewards. Furthermore, the pre-trained policy and mutual

information discriminator can be used to accelerate learning to achieve high tas
k rewards. Our results show that the mutual information between the context stat
es and the states of interest can be an effective ingredient for overcoming chal
lenges in robotic manipulation tasks with sparse rewards. A video showing experi
mental results is available at https://youtu.be/cLRrkd3Y7vU
****

Stagnant zone segmentation with U-net
Selam Waktola,Laurent Babout,Krzysztof Grudzien
Silo discharging and monitoring the process for
industrial or research application depend on computerized
segmentation of different parts of images such as stagnant and
flowing zones which is the toughest task. X-ray Computed
Tomography (CT) is one of a powerful non-destructive technique
for cross-sectional images of a 3D object based on X-ray
absorption. CT is the most proficient for investigating different
granular flow phenomena and segmentation of the stagnant zone
as compared to other imaging techniques. In any case, manual
segmentation is tiresome and erroneous for further investigations.
Hence, automatic and precise strategies are required. In the
present work, a U-net architecture is used for segmenting the
stagnant zone during silo discharging process. This proposed
image segmentation method provides fast and effective outcomes
by exploiting a convolutional neural networks technique with an
accuracy of 97 percent
****

Distance-Based Learning from Errors for Confidence Calibration
Chen Xing,Sercan Arik,Zizhao Zhang,Tomas Pfister
Deep neural networks (DNNs) are poorly calibrated when trained in conventional w
ays. To improve confidence calibration of DNNs, we propose a novel training meth
od, distance-based learning from errors (DBLE). DBLE bases its confidence estima
tion on distances in the representation space. In DBLE, we first adapt prototypi
cal learning to train classification models. It yields a representation space wh
ere the distance between a test sample and its ground truth class center can cal
ibrate the model's classification performance. At inference, however, these dist
ances are not available due to the lack of ground truth labels. To circumvent th
is by inferring the distance for every test sample, we propose to train a confid
ence model jointly with the classification model. We integrate this into trainin
g by merely learning from mis-classified training samples, which we show to be h
ighly beneficial for effective learning. On multiple datasets and DNN architectu
res, we demonstrate that DBLE outperforms alternative single-model confidence ca
libration approaches. DBLE also achieves comparable performance with computation
ally-expensive ensemble approaches with lower computational cost and lower numbe
r of parameters.
****

Curvature Graph Network
Ze Ye,Kin Sum Liu,Tengfei Ma,Jie Gao,Chao Chen
Graph-structured data is prevalent in many domains. Despite the widely celebrate
d success of deep neural networks, their power in graph-structured data is yet t
o be fully explored. We propose a novel network architecture that incorporates a
dvanced graph structural features. In particular, we leverage discrete graph cur
vature, which measures how the neighborhoods of a pair of nodes are structurally
 related. The curvature of an edge (x, y) defines the distance taken to travel f
rom neighbors of x to neighbors of y, compared with the length of edge (x, y). I
t is a much more descriptive feature compared to previously used features that o
nly focus on node specific attributes or limited topological information such as
 degree. Our curvature graph convolution network outperforms state-of-the-art on
 various synthetic and real-world graphs, especially the larger and denser ones.
****

Learning Algorithmic Solutions to Symbolic Planning Tasks with a Neural Computer
Daniel Tanneberg,Elmar Rueckert,Jan Peters

A key feature of intelligent behavior is the ability to learn abstract strategies that transfer to unfamiliar problems. Therefore, we present a novel architecture, based on memory-augmented networks, that is inspired by the von Neumann and Harvard architectures of modern computers. This architecture enables the learning of abstract algorithmic solutions via Evolution Strategies in a reinforcement learning setting. Applied to Sokoban, sliding block puzzle and robotic manipulation tasks, we show that the architecture can learn algorithmic solutions with strong generalization and abstraction: scaling to arbitrary task configurations and complexities, and being independent of both the data representation and the task domain.
****

Generative Imputation and Stochastic Prediction
Mohammad Kachuee,Kimmo Kärkkäinen,Orpaz Goldstein,Sajad Darabi,Majid Sarrafzadeh
In many machine learning applications, we are faced with incomplete datasets. In the literature, missing data imputation techniques have been mostly concerned with filling missing values. However, the existence of missing values is synonymous with uncertainties not only over the distribution of missing values but also over target class assignments that require careful consideration. In this paper, we propose a simple and effective method for imputing missing features and estimating the distribution of target assignments given incomplete data. In order to make imputations, we train a simple and effective generator network to generate imputations that a discriminator network is tasked to distinguish. Following this, a predictor network is trained using the imputed samples from the generator network to capture the classification uncertainties and make predictions accordingly. The proposed method is evaluated on CIFAR-10 image dataset as well as three real-world tabular classification datasets, under different missingness rates and structures. Our experimental results show the effectiveness of the proposed method in generating imputations as well as providing estimates for the class uncertainties in a classification task when faced with missing values.
****

PROTOTYPE-ASSISTED ADVERSARIAL LEARNING FOR UNSUPERVISED DOMAIN ADAPTATION
Dapeng Hu,Jian Liang*,Qibin Hou,Hanshu Yan,Jiashi Feng
This paper presents a generic framework to tackle the crucial class mismatch problem in unsupervised domain adaptation (UDA) for multi-class distributions. Previous adversarial learning methods condition domain alignment only on pseudo labels, but noisy and inaccurate pseudo labels may perturb the multi-class distribution embedded in probabilistic predictions, hence bringing insufficient alleviation to the latent mismatch problem. Compared with pseudo labels, class prototypes are more accurate and reliable since they summarize over all the instances and are able to represent the inherent semantic distribution shared across domains. Therefore, we propose a novel Prototype-Assisted Adversarial Learning (PAAL) scheme, which incorporates instance probabilistic predictions and class prototypes together to provide reliable indicators for adversarial domain alignment. With the PAAL scheme, we align both the instance feature representations and class prototype representations to alleviate the mismatch among semantically different classes. Also, we exploit the class prototypes as proxy to minimize the within-class variance in the target domain to mitigate the mismatch among semantically similar classes. With these novelties, we constitute a Prototype-Assisted Conditional Domain Adaptation (PACDA) framework which well tackles the class mismatch problem. We demonstrate the good performance and generalization ability of the PAAL scheme and also PACDA framework on two UDA tasks, i.e., object recognition (Office-Home,ImageCLEF-DA, andOffice) and synthetic-to-real semantic segmentation (GTA5→CityscapesandSynthia→Cityscapes).
****

Learning Expensive Coordination: An Event-Based Deep RL Approach
Zhenyu Shi*,Runsheng Yu*,Xinrun Wang*,Rundong Wang,Youzhi Zhang,Hanjiang Lai,Bo An
Existing works in deep Multi-Agent Reinforcement Learning (MARL) mainly focus on coordinating cooperative agents to complete certain tasks jointly. However, in many cases of the real world, agents are self-interested such as employees in a

company and clubs in a league. Therefore, the leader, i.e., the manager of the company or the league, needs to provide bonuses to followers for efficient coordination, which we call expensive coordination. The main difficulties of expensive coordination are that i) the leader has to consider the long-term effect and predict the followers' behaviors when assigning bonuses and ii) the complex interactions between followers make the training process hard to converge, especially when the leader's policy changes with time. In this work, we address this problem through an event-based deep RL approach. Our main contributions are threefold. (1) We model the leader's decision-making process as a semi-Markov Decision Process and propose a novel multi-agent event-based policy gradient to learn the leader's long-term policy. (2) We exploit the leader-follower consistency scheme to design a follower-aware module and a follower-specific attention module to predict the followers' behaviors and make accurate response to their behaviors. (3) We propose an action abstraction-based policy gradient algorithm to reduce the followers' decision space and thus accelerate the training process of followers. Experiments in resource collections, navigation, and the predator-prey game reveal that our approach outperforms the state-of-the-art methods dramatically.

****

Unifying Graph Convolutional Networks as Matrix Factorization

Zhaocheng Liu,Qiang Liu,Haoli Zhang,Jun Zhu

In recent years, substantial progress has been made on graph convolutional networks (GCN). In this paper, for the first time, we theoretically analyze the connections between GCN and matrix factorization (MF), and unify GCN as matrix factorization with co-training and unitization. Moreover, under the guidance of this theoretical analysis, we propose an alternative model to GCN named Co-training and Unitized Matrix Factorization (CUMF). The correctness of our analysis is verified by thorough experiments. The experimental results show that CUMF achieves similar or superior performances compared to GCN. In addition, CUMF inherits the benefits of MF-based methods to naturally support constructing mini-batches, and is more friendly to distributed computing comparing with GCN. The distributed CUMF on semi-supervised node classification significantly outperforms distributed GCN methods. Thus, CUMF greatly benefits large scale and complex real-world applications.

****

Frequency Principle: Fourier Analysis Sheds Light on Deep Neural Networks

Zhi-Qin John Xu,Yaoyu Zhang,Tao Luo,Yanyang Xiao,Zheng Ma

We study the training process of Deep Neural Networks (DNNs) from the Fourier analysis perspective. We demonstrate a very universal Frequency Principle (F-Principle) --- DNNs often fit target functions from low to high frequencies --- on high-dimensional benchmark datasets, such as MNIST/CIFAR10, and deep networks, such as VGG16. This F-Principle of DNNs is opposite to the learning behavior of most conventional iterative numerical schemes (e.g., Jacobi method), which exhibits faster convergence for higher frequencies, for various scientific computing problems. With a naive theory, we illustrate that this F-Principle results from the regularity of the commonly used activation functions. The F-Principle implies an implicit bias that DNNs tend to fit training data by a low-frequency function. This understanding provides an explanation of good generalization of DNNs on most real datasets and bad generalization of DNNs on parity function or randomized dataset.

****

Model-free Learning Control of Nonlinear Stochastic Systems with Stability Guarantee

Minghao Han,Yuan Tian,Lixian Zhang,Jun Wang,Wei Pan

Reinforcement learning (RL) offers a principled way to achieve the optimal cumulative performance index in discrete-time nonlinear stochastic systems, which are modeled as Markov decision processes. Its integration with deep learning techniques has promoted the field of deep RL with an impressive performance in complicated continuous control tasks. However, from a control-theoretic perspective, the first and most important property of a system to be guaranteed is stability. Unfortunately, stability is rarely assured in RL and remains an open question. In

this paper, we propose a stability guaranteed RL framework which simultaneously learns a Lyapunov function along with the controller or policy, both of which are parameterized by deep neural networks, by borrowing the concept of Lyapunov function from control theory. Our framework can not only offer comparable or superior control performance over state-of-the-art RL algorithms, but also construct a Lyapunov function to validate the closed-loop stability. In the simulated experiments, our approach is evaluated on several well-known examples including classic CartPole balancing, 3-dimensional robot control and control of synthetic biology gene regulatory networks. Compared with RL algorithms without stability guarantee, our approach can enable the system to recover to the operating point when interfered by uncertainties such as unseen disturbances and system parametric variations to a certain extent.
****

LAMOL: LAnguage MOdeling for Lifelong Language Learning
Fan-Keng Sun*,Cheng-Hao Ho*,Hung-Yi Lee
Most research on lifelong learning applies to images or games, but not language. We present LAMOL, a simple yet effective method for lifelong language learning (LLL) based on language modeling.
LAMOL replays pseudo-samples of previous tasks while requiring no extra memory or model capacity.
Specifically, LAMOL is a language model that simultaneously learns to solve the tasks and generate training samples.
When the model is trained for a new task, it generates pseudo-samples of previous tasks for training alongside data for the new task.
The results show that LAMOL prevents catastrophic forgetting without any sign of intransigence and can perform five very different language tasks sequentially with only one model.
Overall, LAMOL outperforms previous methods by a considerable margin and is only 2-3% worse than multitasking, which is usually considered the LLL upper bound.
The source code is available at https://github.com/jojotenya/LAMOL.
****

GenDICE: Generalized Offline Estimation of Stationary Values
Ruiyi Zhang*,Bo Dai*,Lihong Li,Dale Schuurmans
An important problem that arises in reinforcement learning and Monte Carlo methods is estimating quantities defined by the stationary distribution of a Markov chain. In many real-world applications, access to the underlying transition operator is limited to a fixed set of data that has already been collected, without additional interaction with the environment being available. We show that consistent estimation remains possible in this scenario, and that effective estimation can still be achieved in important applications. Our approach is based on estimating a ratio that corrects for the discrepancy between the stationary and empirical distributions, derived from fundamental properties of the stationary distribution, and exploiting constraint reformulations based on variational divergence minimization. The resulting algorithm, GenDICE, is straightforward and effective. We prove the consistency of the method under general conditions, provide a detailed error analysis, and demonstrate strong empirical performance on benchmark tasks, including off-line PageRank and off-policy policy evaluation.
****

Deep Audio Prior
Yapeng Tian,Chenliang Xu,Dingzeyu Li
Deep convolutional neural networks are known to specialize in distilling compact and robust prior from a large amount of data. We are interested in applying deep networks in the absence of training dataset. In this paper, we introduce deep audio prior (DAP) which leverages the structure of a network and the temporal information in a single audio file. Specifically, we demonstrate that a randomly-initialized neural network can be used with carefully designed audio prior to tackle challenging audio problems such as universal blind source separation, interactive audio editing, audio texture synthesis, and audio co-separation.

To understand the robustness of the deep audio prior, we construct a benchmark d

ataset Universal-150 for universal sound source separation with a diverse set of sources. We show superior audio results than previous work on both qualitatively and quantitative evaluations. We also perform thorough ablation study to validate our design choices.
****
Make Lead Bias in Your Favor: A Simple and Effective Method for News Summarization
Chenguang Zhu,Ziyi Yang,Robert Gmyr,Michael Zeng,Xuedong Huang
Lead bias is a common phenomenon in news summarization, where early parts of an article often contain the most salient information. While many algorithms exploit this fact in summary generation, it has a detrimental effect on teaching the model to discriminate and extract important information. We propose that the lead bias can be leveraged in a simple and effective way in our favor to pretrain abstractive news summarization models on large-scale unlabelled corpus: predicting the leading sentences using the rest of an article. Via careful data cleaning and filtering, our transformer-based pretrained model without any finetuning achieves remarkable results over various news summarization tasks. With further fine tuning, our model outperforms many competitive baseline models. For example, the pretrained model without finetuning outperforms pointer-generator network on CNN/DailyMail dataset. The finetuned model obtains 3.2% higher ROUGE-1, 1.6% higher ROUGE-2 and 2.1% higher ROUGE-L scores than the best baseline model on XSum dataset.
****
ProxSGD: Training Structured Neural Networks under Regularization and Constraints
Yang Yang,Yaxiong Yuan,Avraam Chatzimichailidis,Ruud JG van Sloun,Lei Lei,Symeon Chatzinotas
In this paper, we consider the problem of training neural networks (NN). To promote a NN with specific structures, we explicitly take into consideration the non smooth regularization (such as L1-norm) and constraints (such as interval constraint). This is formulated as a constrained nonsmooth nonconvex optimization problem, and we propose a convergent proximal-type stochastic gradient descent (Prox-SGD) algorithm. We show that under properly selected learning rates, momentum eventually resembles the unknown real gradient and thus is crucial in analyzing the convergence. We establish that with probability 1, every limit point of the sequence generated by the proposed Prox-SGD is a stationary point. Then the Prox-SGD is tailored to train a sparse neural network and a binary neural network, and the theoretical analysis is also supported by extensive numerical tests.
****
Unsupervised Learning of Node Embeddings by Detecting Communities
Chi Thang Duong,Dung Hoang,Truong Giang Le Ba,Thanh Le Cong,Hongzhi Yin,Matthias Weidlich,Quoc Viet Hung Nguyen,Karl Aberer
We present Deep MinCut (DMC), an unsupervised approach to learn node embeddings for graph-structured data. It derives node representations based on their membership in communities. As such, the embeddings directly provide interesting insights into the graph structure, so that the separate node clustering step of existing methods is no longer needed. DMC learns both, node embeddings and communities, simultaneously by minimizing the mincut loss, which captures the number of connections between communities. Striving for high scalability, we also propose a training process for DMC based on minibatches. We provide empirical evidence that the communities learned by DMC are meaningful and that the node embeddings are competitive in different node classification benchmarks.
****
Diverse Trajectory Forecasting with Determinantal Point Processes
Ye Yuan,Kris M. Kitani
The ability to forecast a set of likely yet diverse possible future behaviors of an agent (e.g., future trajectories of a pedestrian) is essential for safety-critical perception systems (e.g., autonomous vehicles). In particular, a set of possible future behaviors generated by the system must be diverse to account for all possible outcomes in order to take necessary safety precautions. It is not s

ufficient to maintain a set of the most likely future outcomes because the set may only contain perturbations of a dominating single outcome (major mode). While generative models such as variational autoencoders (VAEs) have been shown to be a powerful tool for learning a distribution over future trajectories, randomly drawn samples from the learned implicit likelihood model may not be diverse -- the likelihood model is derived from the training data distribution and the samples will concentrate around the major mode of the data. In this work, we propose to learn a diversity sampling function (DSF) that generates a diverse yet likely set of future trajectories. The DSF maps forecasting context features to a set of latent codes which can be decoded by a generative model (e.g., VAE) into a set of diverse trajectory samples. Concretely, the process of identifying the diverse set of samples is posed as DSF parameter estimation. To learn the parameters of the DSF, the diversity of the trajectory samples is evaluated by a diversity loss based on a determinantal point process (DPP). Gradient descent is performed over the DSF parameters, which in turn moves the latent codes of the sample set to find an optimal set of diverse yet likely trajectories. Our method is a novel application of DPPs to optimize a set of items (forecasted trajectories) in continuous space. We demonstrate the diversity of the trajectories produced by our approach on both low-dimensional 2D trajectory data and high-dimensional human motion data.
****
Evaluating The Search Phase of Neural Architecture Search
Kaicheng Yu,Christian Sciuto,Martin Jaggi,Claudiu Musat,Mathieu Salzmann

Neural Architecture Search (NAS) aims to facilitate the design of deep networks for new tasks. Existing techniques rely on two stages: searching over the architecture space and validating the best architecture. NAS algorithms are currently compared solely based on their results on the downstream task. While intuitive, this fails to explicitly evaluate the effectiveness of their search strategies. In this paper, we propose to evaluate the NAS search phase.
To this end, we compare the quality of the solutions obtained by NAS search policies with that of random architecture selection. We find that: (i) On average, the state-of-the-art NAS algorithms perform similarly to the random policy; (ii) the widely-used weight sharing strategy degrades the ranking of the NAS candidates to the point of not reflecting their true performance, thus reducing the effectiveness of the search process.
We believe that our evaluation framework will be key to designing NAS strategies that consistently discover architectures superior to random ones.
****
Learning to Defense by Learning to Attack
Zhehui Chen,Haoming Jiang,Yuyang Shi,Bo Dai,Tuo Zhao
Adversarial training provides a principled approach for training robust neural networks. From an optimization perspective, the adversarial training is essentially solving a minimax robust optimization problem. The outer minimization is trying to learn a robust classifier, while the inner maximization is trying to generate adversarial samples. Unfortunately, such a minimax problem is very difficult to solve due to the lack of convex-concave structure. This work proposes a new adversarial training method based on a generic learning-to-learn (L2L) framework. Specifically, instead of applying the existing hand-designed algorithms for the inner problem, we learn an optimizer, which is parametrized as a convolutional neural network. At the same time, a robust classifier is learned to defense the adversarial attack generated by the learned optimizer. Our experiments over CIFAR-10 and CIFAR-100 datasets demonstrate that the L2L outperforms existing adversarial training methods in both classification accuracy and computational efficiency. Moreover, our L2L framework can be extended to the generative adversarial imitation learning and stabilize the training.
****
On Robustness of Neural Ordinary Differential Equations
Hanshu YAN,Jiawei DU,Vincent TAN,Jiashi FENG
Neural ordinary differential equations (ODEs) have been attracting increasing a

ttention in various research domains recently. There have been some works study ing optimization issues and approximation capabilities of neural ODEs, but their robustness is still yet unclear. In this work, we fill this important gap by ex ploring robustness properties of neural ODEs both empirically and theoretically. We first present an empirical study on the robustness of the neural ODE-based n etworks (ODENets) by exposing them to inputs with various types of perturbations and subsequently investigating the changes of the corresponding outputs. In con trast to conventional convolutional neural networks (CNNs), we find that the ODE Nets are more robust against both random Gaussian perturbations and adversarial attack examples. We then provide an insightful understanding of this phenomenon by exploiting a certain desirable property of the flow of a continuous-time ODE, namely that integral curves are non-intersecting. Our work suggests that, due t o their intrinsic robustness, it is promising to use neural ODEs as a basic bloc k for building robust deep network models. To further enhance the robustness of vanilla neural ODEs, we propose the time-invariant steady neural ODE (TisODE), w hich regularizes the flow on perturbed data via the time-invariant property and the imposition of a steady-state constraint. We show that the TisODE method outp erforms vanilla neural ODEs and also can work in conjunction with other state-of -the-art architectural methods to build more robust deep networks.
****

Diving into Optimization of Topology in Neural Networks
Kun Yuan,Quanquan Li,Yucong Zhou,Jing Shao,Junjie Yan
Seeking effective networks has become one of the most crucial and practical area s in deep learning. The architecture of a neural network can be represented as a directed acyclic graph, whose nodes denote transformation of layers and edges r epresent information flow. Despite the selection of \textit{micro} node operatio ns, \textit{macro} connections among the whole network, noted as \textit{topolog y}, largely affects the optimization process. We first rethink the residual conn ections via a new \textit{topological view} and observe the benefits provided by dense connections to the optimization. Motivated by which, we propose an innova tion method to optimize the topology of a neural network. The optimization space is defined as a complete graph, through assigning learnable weights which refle ct the importance of connections, the optimization of topology is transformed in to learning a set of continuous variables of edges. To extend the optimization to larger search spaces, a new series of networks,
named as TopoNet, are designed. To further focus on critical edges and promote g eneralization ablity in dense topologies, auxiliary sparsity constraint is adopt ed to constrain the distribution of edges. Experiments on classical networks pro ve the effectiveness of the optimization of topology. Experiments with TopoNets further verify both availability and transferability of the proposed method in different tasks e.g. image classification, object detection and face recognition .
****

FoveaBox: Beyound Anchor-based Object Detection
Tao Kong,Fuchun Sun,Huaping Liu,Yuning Jiang,Lei Li,Jianbo Shi
We present FoveaBox, an accurate, flexible, and completely anchor-free framework for object detection. While almost all state-of-the-art object detectors utiliz e predefined anchors to enumerate possible locations, scales and aspect ratios f or the search of the objects, their performance and generalization ability are a lso limited to the design of anchors. Instead, FoveaBox directly learns the obje ct existing possibility and the bounding box coordinates without anchor referenc e. This is achieved by: (a) predicting category-sensitive semantic maps for the object existing possibility, and (b) producing category-agnostic bounding box fo r each position that potentially contains an object. The scales of target boxes are naturally associated with feature pyramid representations. We demonstrate it s effectiveness on standard benchmarks and report extensive experimental analysi s. Without bells and whistles, FoveaBox achieves state-of-the-art single model performance on the standard COCO detection benchmark. More importantly, FoveaBox avoids all computation and hyper-parameters related to anchor boxes, which are often sensitive to the final detection performance. We believe the simple and e

ffective approach will serve as a solid baseline and help ease future research f
or object detection.
****
Cascade Style Transfer
Zhizhong Wang,Lei Zhao,Qihang Mo,Sihuan Lin,Zhiwen Zuo,Wei Xing,Dongming Lu
Recent studies have made tremendous progress in style transfer for specific doma
ins, e.g., artistic, semantic and photo-realistic. However, existing approaches
have limited flexibility in extending to other domains, as different style repre
sentations are often specific to particular domains. This also limits the stylis
tic quality. To address these limitations, we propose Cascade Style Transfer, a
simple yet effective framework that can improve the quality and flexibility of s
tyle transfer by combining multiple existing approaches directly. Our cascade fr
amework contains two architectures, i.e., Serial Style Transfer (SST) and Parall
el Style Transfer (PST). The SST takes the stylized output of one method as the
input content of the others. This could help improve the stylistic quality. The
PST uses a shared backbone and a loss module to optimize the loss functions of d
ifferent methods in parallel. This could help improve the quality and flexibilit
y, and guide us to find domain-independent approaches. Our experiments are condu
cted on three major style transfer domains: artistic, semantic and photo-realist
ic. In all these domains, our methods have shown superiority over the state-of-t
he-art methods.
****
Advantage Weighted Regression: Simple and Scalable Off-Policy Reinforcement Lear
ning
Xue Bin Peng,Aviral Kumar,Grace Zhang,Sergey Levine
In this paper, we aim to develop a simple and scalable reinforcement learning al
gorithm that uses standard supervised learning methods as subroutines. Our goal
is an algorithm that utilizes only simple and convergent maximum likelihood loss
functions, while also being able to leverage off-policy data. Our proposed appr
oach, which we refer to as advantage-weighted regression (AWR), consists of two
standard supervised learning steps: one to regress onto target values for a valu
e function, and another to regress onto weighted target actions for the policy.
The method is simple and general, can accommodate continuous and discrete action
s, and can be implemented in just a few lines of code on top of standard supervi
sed learning methods. We provide a theoretical motivation for AWR and analyze it
s properties when incorporating off-policy data from experience replay. We evalu
ate AWR on a suite of standard OpenAI Gym benchmark tasks, and show that it achi
eves competitive performance compared to a number of well-established state-of-t
he-art RL algorithms. AWR is also able to acquire more effective policies than m
ost off-policy algorithms when learning from purely static datasets with no addi
tional environmental interactions. Furthermore, we demonstrate our algorithm on
challenging continuous control tasks with highly complex simulated characters.
****
Unifying Graph Convolutional Neural Networks and Label Propagation
Hongwei Wang,Jure Leskovec
Label Propagation (LPA) and Graph Convolutional Neural Networks (GCN) are both m
essage passing algorithms on graphs. Both solve the task of node classification
but LPA propagates node label information across the edges of the graph, while G
CN propagates and transforms node feature information. However, while conceptual
ly similar, theoretical relation between LPA and GCN has not yet been investigat
ed. Here we study the relationship between LPA and GCN in terms of two aspects:
(1) feature/label smoothing where we analyze how the feature/label of one node a
re spread over its neighbors; And, (2) feature/label influence of how much the i
nitial feature/label of one node influences the final feature/label of another n
ode. Based on our theoretical analysis, we propose an end-to-end model that unif
ies GCN and LPA for node classification. In our unified model, edge weights are
learnable, and the LPA serves as regularization to assist the GCN in learning pr
oper edge weights that lead to improved classification performance. Our model ca
n also be seen as learning attention weights based on node labels, which is more
task-oriented than existing feature-based attention models. In a number of expe

riments on real-world graphs, our model shows superiority over state-of-the-art GCN-based methods in terms of node classification accuracy.

****
Equivariant neural networks and equivarification
Erkao Bao,Linqi Song
A key difference from existing works is that our equivarification method can be applied without knowledge of the detailed functions of a layer in a neural network, and hence, can be generalized to any feedforward neural networks. Although the network size scales up, the constructed equivariant neural network does not increase the complexity of the network compared with the original one, in terms of the number of parameters. As an illustration, we build an equivariant neural network for image classification by equivarifying a convolutional neural network. Results show that our proposed method significantly reduces the design and training complexity, yet preserving the learning performance in terms of accuracy.
****
Data Valuation using Reinforcement Learning
Jinsung Yoon,Sercan O. Arik,Tomas Pfister
Quantifying the value of data is a fundamental problem in machine learning. Data valuation has multiple important use cases: (1) building insights about the learning task, (2) domain adaptation, (3) corrupted sample discovery, and (4) robust learning. To adaptively learn data values jointly with the target task predictor model, we propose a meta learning framework which we name Data Valuation using Reinforcement Learning (DVRL). We employ a data value estimator (modeled by a deep neural network) to learn how likely each datum is used in training of the predictor model. We train the data value estimator using a reinforcement signal of the reward obtained on a small validation set that reflects performance on the target task. We demonstrate that DVRL yields superior data value estimates compared to alternative methods across different types of datasets and in a diverse set of application scenarios. The corrupted sample discovery performance of DVRL is close to optimal in many regimes (i.e. as if the noisy samples were known apriori), and for domain adaptation and robust learning DVRL significantly outperforms state-of-the-art by 14.6% and 10.8%, respectively.
****
RL-LIM: Reinforcement Learning-based Locally Interpretable Modeling
Jinsung Yoon,Sercan O. Arik,Tomas Pfister
Understanding black-box machine learning models is important towards their widespread adoption. However, developing globally interpretable models that explain the behavior of the entire model is challenging. An alternative approach is to explain black-box models through explaining individual prediction using a locally interpretable model. In this paper, we propose a novel method for locally interpretable modeling -- Reinforcement Learning-based Locally Interpretable Modeling (RL-LIM). RL-LIM employs reinforcement learning to select a small number of samples and distill the black-box model prediction into a low-capacity locally interpretable model. Training is guided with a reward that is obtained directly by measuring agreement of the predictions from the locally interpretable model with the black-box model. RL-LIM near-matches the overall prediction performance of black-box models while yielding human-like interpretability, and significantly outperforms state of the art locally interpretable models in terms of overall prediction performance and fidelity.
****
BackPACK: Packing more into Backprop
Felix Dangel,Frederik Kunstner,Philipp Hennig
Automatic differentiation frameworks are optimized for exactly one thing: computing the average mini-batch gradient. Yet, other quantities such as the variance of the mini-batch gradients or many approximations to the Hessian can, in theory, be computed efficiently, and at the same time as the gradient. While these quantities are of great interest to researchers and practitioners, current deep learning software does not support their automatic calculation. Manually implementing them is burdensome, inefficient if done naively, and the resulting co

de is rarely shared. This hampers progress in deep learning, and unnecessa rily narrows research to focus on gradient descent and its variants; it also complicates replication studies and comparisons between newly develop ed methods that require those quantities, to the point of impossibility. To addr ess this problem, we introduce BackPACK, an efficient framework built on top of PyTorch, that extends the backpropagation algorithm to extract additional info rmation from first-and second-order derivatives. Its capabilities are illustrate d by benchmark reports for computing additional quantities on deep neural networ ks, and an example application by testing several recent curvature approximation s for optimization.
****

DeepHoyer: Learning Sparser Neural Network with Differentiable Scale-Invariant S parsity Measures
Huanrui Yang,Wei Wen,Hai Li
In seeking for sparse and efficient neural network models, many previous works i nvestigated on enforcing L1 or L0 regularizers to encourage weight sparsity duri ng training. The L0 regularizer measures the parameter sparsity directly and is invariant to the scaling of parameter values. But it cannot provide useful gradi ents and therefore requires complex optimization techniques. The L1 regularizer is almost everywhere differentiable and can be easily optimized with gradient de scent. Yet it is not scale-invariant and causes the same shrinking rate to all p arameters, which is inefficient in increasing sparsity. Inspired by the Hoyer me asure (the ratio between L1 and L2 norms) used in traditional compressed sensing problems, we present DeepHoyer, a set of sparsity-inducing regularizers that ar e both differentiable almost everywhere and scale-invariant. Our experiments sho w that enforcing DeepHoyer regularizers can produce even sparser neural network models than previous works, under the same accuracy level. We also show that Dee pHoyer can be applied to both element-wise and structural pruning.
****

Regional based query in graph active learning
Abel Roy,Louzoun Yoram
Graph convolution networks (GCN) have emerged as a leading method to classify no des and graphs. These GCN have been combined with active learning (AL) methods, when a small chosen set of tagged examples can be used. Most AL-GCN use the sam ple class uncertainty as selection criteria, and not the graph. In contrast, rep resentative sampling uses the graph, but not the prediction. We propose to combi ne the two and query nodes based on the uncertainty of the graph around them. We here propose two novel methods to select optimal nodes in AL-GCN that explicitl y use the graph information to query for optimal nodes. The first method named r egional uncertainty is an extension of the classical entropy measure, but instea d of sampling nodes with high entropy, we propose to sample nodes surrounded by nodes of different classes, or nodes with high ambiguity. The second method call ed Adaptive Page-Rank is an extension of the page-rank algorithm, where nodes t hat have a low probability of being reached by random walks from tagged nodes ar e selected. We show that the latter is optimal when the fraction of tagged nodes is low, and when this fraction grows to one over the average degree, the region al uncertainty performs better than all existing methods. While we have tested t hese methods on graphs, such methods can be extended to any classification probl em, where a distance can be defined between the input samples.
****

Group-Connected Multilayer Perceptron Networks
Mohammad Kachuee,Sajad Darabi,Shayan Fazeli,Majid Sarrafzadeh
Despite the success of deep learning in domains such as image, voice, and graphs , there has been little progress in deep representation learning for domains wit hout a known structure between features. For instance, a tabular dataset of diff erent demographic and clinical factors where the feature interactions are not gi ven as a prior. In this paper, we propose Group-Connected Multilayer Perceptron (GMLP) networks to enable deep representation learning in these domains. GMLP is based on the idea of learning expressive feature combinations (groups) and expl oiting them to reduce the network complexity by defining local group-wise operat

ions. During the training phase, GMLP learns a sparse feature grouping matrix us ing temperature annealing softmax with an added entropy loss term to encourage t he sparsity. Furthermore, an architecture is suggested which resembles binary tr ees, where group-wise operations are followed by pooling operations to combine i nformation; reducing the number of groups as the network grows in depth. To eval uate the proposed method, we conducted experiments on five different real-world datasets covering various application areas. Additionally, we provide visualizat ions on MNIST and synthesized data. According to the results, GMLP is able to su ccessfully learn and exploit expressive feature combinations and achieve state-o f-the-art classification performance on different datasets.
****

Towards Stable and comprehensive Domain Alignment: Max-Margin Domain-Adversarial Training
Jianfei Yang,Han Zou,Yuxun Zhou,Lihua Xie
Domain adaptation tackles the problem of transferring knowledge from a label-r ich source domain to an unlabeled or label-scarce target domain. Recently domain -adversarial training (DAT) has shown promising capacity to learn a domain-invar iant feature space by reversing the gradient propagation of a domain classifier. However, DAT is still vulnerable in several aspects including (1) training inst ability due to the overwhelming discriminative ability of the domain classifier in adversarial training, (2) restrictive feature-level alignment, and (3) lack o f interpretability or systematic explanation of the learned feature space. In th is paper, we propose a novel Max-margin Domain-Adversarial Training (MDAT) by de signing an Adversarial Reconstruction Network (ARN). The proposed MDAT stabilize s the gradient reversing in ARN by replacing the domain classifier with a recons truction network, and in this manner ARN conducts both feature-level and pixel-l evel domain alignment without involving extra network structures. Furthermore, A RN demonstrates strong robustness to a wide range of hyper-parameters settings, greatly alleviating the task of model selection. Extensive empirical results val idate that our approach outperforms other state-of-the-art domain alignment meth ods. Additionally, the reconstructed target samples are visualized to interpret the domain-invariant feature space which conforms with our intuition.
****

Depth-Adaptive Transformer
Maha Elbayad,Jiatao Gu,Edouard Grave,Michael Auli
State of the art sequence-to-sequence models for large scale tasks perform a fix ed number of computations for each input sequence regardless of whether it is ea sy or hard to process. In this paper, we train Transformer models which can make output predictions at different stages of the network and we investigate differ ent ways to predict how much computation is required for a particular sequence. Unlike dynamic computation in Universal Transformers, which applies the same set of layers iteratively, we apply different layers at every step to adjust both t he amount of computation as well as the model capacity. On IWSLT German-English translation our approach matches the accuracy of a well tuned baseline Transform er while using less than a quarter of the decoder layers.
****

InfoGraph: Unsupervised and Semi-supervised Graph-Level Representation Learning via Mutual Information Maximization
Fan-Yun Sun,Jordan Hoffman,Vikas Verma,Jian Tang
This paper studies learning the representations of whole graphs in both unsuperv ised and semi-supervised scenarios. Graph-level representations are critical in a variety of real-world applications such as predicting the properties of molecu les and community analysis in social networks. Traditional graph kernel based me thods are simple, yet effective for obtaining fixed-length representations for g raphs but they suffer from poor generalization due to hand-crafted designs. Ther e are also some recent methods based on language models (e.g. graph2vec) but the y tend to only consider certain substructures (e.g. subtrees) as graph represent atives. Inspired by recent progress of unsupervised representation learning, in this paper we proposed a novel method called InfoGraph for learning graph-level representations. We maximize the mutual information between the graph-level repr

esentation and the representations of substructures of different scales (e.g., n
odes, edges, triangles). By doing so, the graph-level representations encode asp
ects of the data that are shared across different scales of substructures. Furth
ermore, we further propose InfoGraph*, an extension of InfoGraph for semisupervi
sed scenarios. InfoGraph* maximizes the mutual information between unsupervised
graph representations learned by InfoGraph and the representations learned by ex
isting supervised methods. As a result, the supervised encoder learns from unlab
eled data while preserving the latent semantic space favored by the current supe
rvised task. Experimental results on the tasks of graph classification and molec
ular property prediction show that InfoGraph is superior to state-of-the-art bas
elines and InfoGraph* can achieve performance competitive with state-of-the-art
semi-supervised models.
****

Federated Adversarial Domain Adaptation
Xingchao Peng,Zijun Huang,Yizhe Zhu,Kate Saenko
Federated learning improves data privacy and efficiency in machine learning perf
ormed over networks of distributed devices, such as mobile phones, IoT and weara
ble devices, etc. Yet models trained with federated learning can still fail to g
eneralize to new devices due to the problem of domain shift. Domain shift occurs
 when the labeled data collected by source nodes statistically differs from the
target node's unlabeled data. In this work, we present a principled approach to
the problem of federated domain adaptation, which aims to align the representati
ons learned among the different nodes with the data distribution of the target n
ode. Our approach extends adversarial adaptation techniques to the constraints o
f the federated setting. In addition, we devise a dynamic attention mechanism an
d leverage feature disentanglement to enhance knowledge transfer. Empirically, w
e perform extensive experiments on several image and text classification tasks a
nd show promising results under unsupervised federated domain adaptation setting
.
****

CATER: A diagnostic dataset for Compositional Actions & TEmporal Reasoning
Rohit Girdhar,Deva Ramanan
Computer vision has undergone a dramatic revolution in performance, driven in la
rge part through deep features trained on large-scale supervised datasets. Howev
er, much of these improvements have focused on static image analysis; video unde
rstanding has seen rather modest improvements. Even though new datasets and spat
iotemporal models have been proposed, simple frame-by-frame classification metho
ds often still remain competitive. We posit that current video datasets are plag
ued with implicit biases over scene and object structure that can dwarf variatio
ns in temporal structure. In this work, we build a video dataset with fully obse
rvable and controllable object and scene bias, and which truly requires spatiote
mporal understanding in order to be solved. Our dataset, named CATER, is rendere
d synthetically using a library of standard 3D objects, and tests the ability to
 recognize compositions of object movements that require long-term reasoning. In
 addition to being a challenging dataset, CATER also provides a plethora of diag
nostic tools to analyze modern spatiotemporal video architectures by being compl
etely observable and controllable. Using CATER, we provide insights into some of
 the most recent state of the art deep video architectures.
****

Learning Structured Communication for Multi-agent Reinforcement Learning
Junjie Sheng,Xiangfeng Wang,Bo Jin,Junchi Yan,Wenhao Li,Tsung-Hui Chang,Jun Wang
,Hongyuan Zha
Learning to cooperate is crucial for many practical large-scale multi-agent appl
ications. In this work, we consider an important collaborative task, in which ag
ents learn to ef■ciently communicate with each other under a multi-agent reinfor
cement learning (MARL) setting. Despite the fact that there has been a number of
 existing works along this line, achieving global cooperation at scale is still
challenging. In particular, most of the existing algorithms suffer from issues s
uch as scalability and high communication complexity, in the sense that when the
 agent population is large, it can be dif■cult to extract effective information

for high-performance MARL. In contrast, the proposed algorithmic framework, term ed Learning Structured Communication (LSC), is not only scalable but also commun ication high-qualitative (learning ef■cient). The key idea is to allow the agent s to dynamically learn a hierarchical communication structure, while under such a structure the graph neural network (GNN) is used to ef■ciently extract useful information to be exchanged between the neighboring agents. A number of new tech niques are proposed to tightly integrate the communication structure learning, G NN optimization and MARL tasks. Extensive experiments are performed to demonstra te that, the proposed LSC framework enjoys high communication ef■ciency, scalabi lity and global cooperation capability.
****

Utilizing Edge Features in Graph Neural Networks via Variational Information Max imization

Pengfei Chen,Weiwen Liu,Chang-Yu Hsieh,Guangyong Chen,Pheng Ann Heng

Graph Neural Networks (GNNs) broadly follow the scheme that the representation v ector of each node is updated recursively using the message from neighbor nodes, where the message of a neighbor is usually pre-processed with a parameterized t ransform matrix. To make better use of edge features, we propose the Edge Inform ation maximized Graph Neural Network (EIGNN) that maximizes the Mutual Informati on (MI) between edge features and message passing channels. The MI is reformulat ed as a differentiable objective via a variational approach. We theoretically sh ow that the newly introduced objective enables the model to preserve edge inform ation, and empirically corroborate the enhanced performance of MI-maximized mode ls across a broad range of learning tasks including regression on molecular grap hs and relation prediction in knowledge graphs.
****

Stabilizing DARTS with Amended Gradient Estimation on Architectural Parameters

Kaifeng Bi,Changping Hu,Lingxi Xie,Xin Chen,Longhui Wei,Qi Tian

Differentiable neural architecture search has been a popular methodology of expl oring architectures for deep learning. Despite the great advantage of search eff iciency, it often suffers weak stability, which obstacles it from being applied to a large search space or being flexibly adjusted to different scenarios. This paper investigates DARTS, the currently most popular differentiable search algor ithm, and points out an important factor of instability, which lies in its appro ximation on the gradients of architectural parameters. In the current status, th e optimization algorithm can converge to another point which results in dramatic inaccuracy in the re-training process. Based on this analysis, we propose an am ending term for computing architectural gradients by making use of a direct prop erty of the optimality of network parameter optimization. Our approach mathemati cally guarantees that gradient estimation follows a roughly correct direction, w hich leads the search stage to converge on reasonable architectures. In practice , our algorithm is easily implemented and added to DARTS-based approaches effici ently. Experiments on CIFAR and ImageNet demonstrate that our approach enjoys ac curacy gain and, more importantly, enables DARTS-based approaches to explore muc h larger search spaces that have not been studied before.
****

Effective Mechanism to Mitigate Injuries During NFL Plays

Arraamuthan Arulanantham,Ahamed Arshad Ahamed Anzar,Gowshalini Rajalingam,Krusan th Ingran,Prasanna S. Haddela

NFL(American football),which is regarded as the premier sports icon of America, has been severely accused in the recent years of being exposed to dangerous inju ries that prove to be a bigger crisis as the players' lives have been increasing ly at risk. Concussions, which refer to the serious brain traumas experienced du ring the passage of NFL play, have displayed a dramatic rise in the recent seas ons concluding in an alarming rate in 2017/18. Acknowledging the potential risk, the NFL has been trying to fight via NeuroIntel AI mechanism as well as modifyi ng existing game rules and risky play practices to reduce the rate of concussio ns. As a remedy, we are suggesting an effective mechanism to extensively analyse the potential concussion risks by adopting predictive analysis to project injur y risk percentage per each play and positional impact analysis to suggest safer

team formation pairs to lessen injuries to offer a comprehensive study on NFL injury analysis. The proposed data analytical approach differentiates itself from the other similar approaches that were focused only on the descriptive analysis rather than going for a bigger context with predictive modelling and formation pairs mining that would assist in modifying existing rules to tackle injury concerns. The predictive model that works with Kafka-stream processor real-time inputs and risky formation pairs identification by designing FP-Matrix, makes this far-reaching solution to analyse injury data on various grounds wherever applicable.
****

TechKG: A Large-Scale Chinese Technology-Oriented Knowledge Graph
Feiliang Ren
Knowledge graph is a kind of valuable knowledge base which would benefit lots of AI-related applications. Up to now, lots of large-scale knowledge graphs have been built. However, most of them are non-Chinese and designed for general purpose. In this work, we introduce TechKG, a large scale Chinese knowledge graph that is technology-oriented. It is built automatically from massive technical papers that are published in Chinese academic journals of different research domains. Some carefully designed heuristic rules are used to extract high quality entities and relations. Totally, it comprises of over 260 million triplets that are built upon more than 52 million entities which come from 38 research domains. Our preliminary experiments indicate that TechKG has high adaptability and can be used as a dataset for many diverse AI-related applications.
****

Learning Reusable Options for Multi-Task Reinforcement Learning
Francisco M. Garcia,Chris Nota,Philip S. Thomas
Reinforcement learning (RL) has become an increasingly active area of research in recent years. Although there are many algorithms that allow an agent to solve tasks efficiently, they often ignore the possibility that prior experience related to the task at hand might be available. For many practical applications, it might be unfeasible for an agent to learn how to solve a task from scratch, given that it is generally a computationally expensive process; however, prior experience could be leveraged to make these problems tractable in practice. In this paper, we propose a framework for exploiting existing experience by learning reusable options. We show that after an agent learns policies for solving a small number of problems, we are able to use the trajectories generated from those policies to learn reusable options that allow an agent to quickly learn how to solve novel and related problems.
****

Maxmin Q-learning: Controlling the Estimation Bias of Q-learning
Qingfeng Lan,Yangchen Pan,Alona Fyshe,Martha White
Q-learning suffers from overestimation bias, because it approximates the maximum action value using the maximum estimated action value. Algorithms have been proposed to reduce overestimation bias, but we lack an understanding of how bias interacts with performance, and the extent to which existing algorithms mitigate bias. In this paper, we 1) highlight that the effect of overestimation bias on learning efficiency is environment-dependent; 2) propose a generalization of Q-learning, called \emph{Maxmin Q-learning}, which provides a parameter to flexibly control bias; 3) show theoretically that there exists a parameter choice for Maxmin Q-learning that leads to unbiased estimation with a lower approximation variance than Q-learning; and 4) prove the convergence of our algorithm in the tabular case, as well as convergence of several previous Q-learning variants, using a novel Generalized Q-learning framework. We empirically verify that our algorithm better controls estimation bias in toy environments, and that it achieves superior performance on several benchmark problems.
****

X-Forest: Approximate Random Projection Trees for Similarity Measurement
Yikai Zhao,Peiqing Chen,Zidong Zhao,Tong Yang,Jie Jiang,Bin Cui,Gong Zhang,Steve Uhlig
Similarity measurement plays a central role in various data mining and machine l

earning tasks. Generally, a similarity measurement solution should, in an ideal state, possess the following three properties: accuracy, efficiency and independence from prior knowledge. Yet unfortunately, vital as similarity measurements are, no previous works have addressed all of them. In this paper, we propose X-Forest, consisting of a group of approximate Random Projection Trees, such that all three targets mentioned above are tackled simultaneously. Our key techniques are as follows. First, we introduced RP Trees into the tasks of similarity measurement such that accuracy is improved. In addition, we enforce certain layers in each tree to share identical projection vectors, such that exalted efficiency is achieved. Last but not least, we introduce randomness into partition to eliminate its reliance on prior knowledge. We conduct experiments on three real-world datasets, whose results demonstrate that our model, X-Forest, reaches an efficiency of up to 3.5 times higher than RP Trees with negligible compromising on its accuracy, while also being able to outperform traditional Euclidean distance-based similarity metrics by as much as 20% with respect to clustering tasks. We have released codes in github anonymously so as to meet the demand of reproducibility.
****

Low Bias Gradient Estimates for Very Deep Boolean Stochastic Networks
Adeel Pervez,Taco Cohen,Efstratios Gavves
Stochastic neural networks with discrete random variables are an important class of models for their expressivity and interpretability. Since direct differentiation and backpropagation is not possible, Monte Carlo gradient estimation techniques have been widely employed for training such models. Efficient stochastic gradient estimators, such Straight-Through and Gumbel-Softmax, work well for shallow models with one or two stochastic layers. Their performance, however, suffers with increasing model complexity.
In this work we focus on stochastic networks with multiple layers of Boolean latent variables. To analyze such such networks, we employ the framework of harmonic analysis for Boolean functions. We use it to derive an analytic formulation for the source of bias in the biased Straight-Through estimator. Based on the analysis we propose \emph{FouST}, a simple gradient estimation algorithm that relies on three simple bias reduction steps. Extensive experiments show that FouST performs favorably compared to state-of-the-art biased estimators, while being much faster than unbiased ones. To the best of our knowledge FouST is the first gradient estimator to train up very deep stochastic neural networks, with up to 80 deterministic and 11 stochastic layers.
****

Automatically Discovering and Learning New Visual Categories with Ranking Statistics
Kai Han,Sylvestre-Alvise Rebuffi,Sebastien Ehrhardt,Andrea Vedaldi,Andrew Zisserman
We tackle the problem of discovering novel classes in an image collection given labelled examples of other classes. This setting is similar to semi-supervised learning, but significantly harder because there are no labelled examples for the new classes. The challenge, then, is to leverage the information contained in the labelled images in order to learn a general-purpose clustering model and use the latter to identify the new classes in the unlabelled data. In this work we address this problem by combining three ideas: (1) we suggest that the common approach of bootstrapping an image representation using the labeled data only introduces an unwanted bias, and that this can be avoided by using self-supervised learning to train the representation from scratch on the union of labelled and unlabelled data; (2) we use rank statistics to transfer the model's knowledge of the labelled classes to the problem of clustering the unlabelled images; and, (3) we train the data representation by optimizing a joint objective function on the labelled and unlabelled subsets of the data, improving both the supervised classification of the labelled data, and the clustering of the unlabelled data. We evaluate our approach on standard classification benchmarks and outperform current methods for novel category discovery by a significant margin.
****

Support-guided Adversarial Imitation Learning
Ruohan Wang,Carlo Ciliberto,Pierluigi Amadori,Yiannis Demiris
We propose Support-guided Adversarial Imitation Learning (SAIL), a generic imita tion learning framework that unifies support estimation of the expert policy wit h the family of Adversarial Imitation Learning (AIL) algorithms. SAIL addresses two important challenges of AIL, including the implicit reward bias and potentia l training instability. We also show that SAIL is at least as efficient as stand ard AIL. In an extensive evaluation, we demonstrate that the proposed method eff ectively handles the reward bias and achieves better performance and training st ability than other baseline methods on a wide range of benchmark control tasks.
****
Mutual Mean-Teaching: Pseudo Label Refinery for Unsupervised Domain Adaptation o n Person Re-identification
Yixiao Ge,Dapeng Chen,Hongsheng Li
Person re-identification (re-ID) aims at identifying the same persons' images ac ross different cameras. However, domain diversities between different datasets p ose an evident challenge for adapting the re-ID model trained on one dataset to another one. State-of-the-art unsupervised domain adaptation methods for person re-ID transferred the learned knowledge from the source domain by optimizing wit h pseudo labels created by clustering algorithms on the target domain. Although they achieved state-of-the-art performances, the inevitable label noise caused b y the clustering procedure was ignored. Such noisy pseudo labels substantially h inders the model's capability on further improving feature representations on th e target domain. In order to mitigate the effects of noisy pseudo labels, we pro pose to softly refine the pseudo labels in the target domain by proposing an uns upervised framework, Mutual Mean-Teaching (MMT), to learn better features from t he target domain via off-line refined hard pseudo labels and on-line refined sof t pseudo labels in an alternative training manner. In addition, the common prac tice is to adopt both the classification loss and the triplet loss jointly for a chieving optimal performances in person re-ID models. However, conventional trip let loss cannot work with softly refined labels. To solve this problem, a novel soft softmax-triplet loss is proposed to support learning with soft pseudo tripl et labels for achieving the optimal domain adaptation performance. The proposed MMT framework achieves considerable improvements of 14.4%, 18.2%, 13.1% and 16.4 % mAP on Market-to-Duke, Duke-to-Market, Market-to-MSMT and Duke-to-MSMT unsuper vised domain adaptation tasks.
****
Multi-Scale Representation Learning for Spatial Feature Distributions using Gri d Cells
Gengchen Mai,Krzysztof Janowicz,Bo Yan,Rui Zhu,Ling Cai,Ni Lao
Unsupervised text encoding models have recently fueled substantial progress in N LP. The key idea is to use neural networks to convert words in texts to vector s pace representations (embeddings) based on word positions in a sentence and thei r contexts, which are suitable for end-to-end training of downstream tasks. We s ee a strikingly similar situation in spatial analysis, which focuses on incorpor ating both absolute positions and spatial contexts of geographic objects such as POIs into models. A general-purpose representation model for space is valuable for a multitude of tasks. However, no such general model exists to date beyond s imply applying discretization or feed-forward nets to coordinates, and little ef fort has been put into jointly modeling distributions with vastly different char acteristics, which commonly emerges from GIS data. Meanwhile, Nobel Prize-winnin g Neuroscience research shows that grid cells in mammals provide a multi-scale p eriodic representation that functions as a metric for location encoding and is c ritical for recognizing places and for path-integration. Therefore, we propose a representation learning model called Space2Vec to encode the absolute positions and spatial relationships of places. We conduct experiments on two real-world g eographic data for two different tasks: 1) predicting types of POIs given their positions and context, 2) image classification leveraging their geo-locations. R esults show that because of its multi-scale representations, Space2Vec outperfor ms well-established ML approaches such as RBF kernels, multi-layer feed-forward

nets, and tile embedding approaches for location modeling and image classificati
on tasks. Detailed analysis shows that all baselines can at most well handle dis
tribution at one scale but show poor performances in other scales. In contrast,
Space2Vec ’s multi-scale representation can handle distributions at different sc
ales.
****

Neural Oblivious Decision Ensembles for Deep Learning on Tabular Data
Sergei Popov,Stanislav Morozov,Artem Babenko
Nowadays, deep neural networks (DNNs) have become the main instrument for machin
e learning tasks within a wide range of domains, including vision, NLP, and spee
ch. Meanwhile, in an important case of heterogenous tabular data, the advantage
of DNNs over shallow counterparts remains questionable. In particular, there is
no sufficient evidence that deep learning machinery allows constructing methods
that outperform gradient boosting decision trees (GBDT), which are often the top
 choice for tabular problems. In this paper, we introduce Neural Oblivious Decis
ion Ensembles (NODE), a new deep learning architecture, designed to work with an
y tabular data. In a nutshell, the proposed NODE architecture generalizes ensemb
les of oblivious decision trees, but benefits from both end-to-end gradient-base
d optimization and the power of multi-layer hierarchical representation learning
. With an extensive experimental comparison to the leading GBDT packages on a la
rge number of tabular datasets, we demonstrate the advantage of the proposed NOD
E architecture, which outperforms the competitors on most of the tasks. We open-
source the PyTorch implementation of NODE and believe that it will become a univ
ersal framework for machine learning on tabular data.
****

Data augmentation instead of explicit regularization
Alex Hernandez-Garcia,Peter König
Modern deep artificial neural networks have achieved impressive results through
models with orders of magnitude more parameters than training examples which con
trol overfitting with the help of regularization. Regularization can be implicit
, as is the case of stochastic gradient descent and parameter sharing in convolu
tional layers, or explicit. Explicit regularization techniques, most common form
s are weight decay and dropout, have proven successful in terms of improved gene
ralization, but they blindly reduce the effective capacity of the model, introdu
ce sensitive hyper-parameters and require deeper and wider architectures to comp
ensate for the reduced capacity. In contrast, data augmentation techniques explo
it domain knowledge to increase the number of training examples and improve gene
ralization without reducing the effective capacity and without introducing model
-dependent parameters, since it is applied on the training data. In this paper w
e systematically contrast data augmentation and explicit regularization on three
 popular architectures and three data sets. Our results demonstrate that data au
gmentation alone can achieve the same performance or higher as regularized model
s and exhibits much higher adaptability to changes in the architecture and the a
mount of training data.
****

SQIL: Imitation Learning via Reinforcement Learning with Sparse Rewards
Siddharth Reddy,Anca D. Dragan,Sergey Levine
Learning to imitate expert behavior from demonstrations can be challenging, espe
cially in environments with high-dimensional, continuous observations and unknow
n dynamics. Supervised learning methods based on behavioral cloning (BC) suffer
from distribution shift: because the agent greedily imitates demonstrated action
s, it can drift away from demonstrated states due to error accumulation. Recent
methods based on reinforcement learning (RL), such as inverse RL and generative
adversarial imitation learning (GAIL), overcome this issue by training an RL age
nt to match the demonstrations over a long horizon. Since the true reward functi
on for the task is unknown, these methods learn a reward function from the demon
strations, often using complex and brittle approximation techniques that involve
 adversarial training. We propose a simple alternative that still uses RL, but d
oes not require learning a reward function. The key idea is to provide the agent
 with an incentive to match the demonstrations over a long horizon, by encouragi

ng it to return to demonstrated states upon encountering new, out-of-distributio n states. We accomplish this by giving the agent a constant reward of r=+1 for m atching the demonstrated action in a demonstrated state, and a constant reward o f r=0 for all other behavior. Our method, which we call soft Q imitation learnin g (SQIL), can be implemented with a handful of minor modifications to any standa rd Q-learning or off-policy actor-critic algorithm. Theoretically, we show that SQIL can be interpreted as a regularized variant of BC that uses a sparsity prio r to encourage long-horizon imitation. Empirically, we show that SQIL outperform s BC and achieves competitive results compared to GAIL, on a variety of image-ba sed and low-dimensional tasks in Box2D, Atari, and MuJoCo. This paper is a proof of concept that illustrates how a simple imitation method based on RL with cons tant rewards can be as effective as more complex methods that use learned reward s.

****

Label Cleaning with Likelihood Ratio Test

Songzhu Zheng,Pengxiang Wu,Aman Goswami,Mayank Goswami,Dimitris Metaxas,Chao Che n

To collect large scale annotated data, it is inevitable to introduce label noise , i.e., incorrect class labels. A major challenge is to develop robust deep lear ning models that achieve high test performance despite training set label noise. We introduce a novel approach that directly cleans labels in order to train a high quality model. Our method leverages statistical principles to correct data labels and has a theoretical guarantee of the correctness. In particular, we us e a likelihood ratio test(LRT) to flip the labels of training data. We prove th at our LRT label correction algorithm is guaranteed to flip the label so it is c onsistent with the true Bayesian optimal decision rule with high probability. W e incorporate our label correction algorithm into the training of deep neural ne tworks and train models that achieve superior testing performance on multiple pu blic datasets.

****

Visual Imitation with Reinforcement Learning using Recurrent Siamese Networks

Glen Berseth,Christopher Pal

It would be desirable for a reinforcement learning (RL) based agent to learn beh aviour by merely watching a demonstration. However, defining rewards that facil itate this goal within the RL paradigm remains a challenge. Here we address this problem with Siamese networks, trained to compute distances between observed be haviours and the agent’s behaviours. Given a desired motion such Siamese network s can be used to provide a reward signal to an RL agent via the distance between the desired motion and the agent’s motion. We experiment with an RNN-based comp arator model that can compute distances in space and time between motion clips w hile training an RL policy to minimize this distance. Through experimentation, we have had also found that the inclusion of multi-task data and an additional i mage encoding loss helps enforce the temporal consistency. These two components appear to balance reward for matching a specific instance of a behaviour versus that behaviour in general. Furthermore, we focus here on a particularly challen ging form of this problem where only a single demonstration is provided for a gi ven task – the one-shot learning setting. We demonstrate our approach on humanoi d agents in both 2D with 10 degrees of freedom (DoF) and 3D with 38 DoF.

****

Graph Neural Networks Exponentially Lose Expressive Power for Node Classificatio n

Kenta Oono,Taiji Suzuki

Graph Neural Networks (graph NNs) are a promising deep learning approach for ana lyzing graph-structured data. However, it is known that they do not improve (or sometimes worsen) their predictive performance as we pile up many layers and add non-lineality. To tackle this problem, we investigate the expressive power of g raph NNs via their asymptotic behaviors as the layer size tends to infinity. Our strategy is to generalize the forward propagation of a Graph Convolutional N etwork (GCN), which is a popular graph NN variant, as a specific dynamical syste m. In the case of a GCN, we show that when its weights satisfy the conditions de

termined by the spectra of the (augmented) normalized Laplacian, its output expo
nentially approaches the set of signals that carry information of the connected
components and node degrees only for distinguishing nodes.
Our theory enables us to relate the expressive power of GCNs with the topologica
l information of the underlying graphs inherent in the graph spectra. To demonst
rate this, we characterize the asymptotic behavior of GCNs on the Erd\H{o}s -- R
\'{e}nyi graph.
We show that when the Erd\H{o}s -- R\'{e}nyi graph is sufficiently dense and lar
ge, a broad range of GCNs on it suffers from the ``information loss" in the limi
t of infinite layers with high probability.
Based on the theory, we provide a principled guideline for weight normalization
of graph NNs. We experimentally confirm that the proposed weight scaling enhance
s the predictive performance of GCNs in real data. Code is available at https://
github.com/delta2323/gnn-asymptotics.
****

VIDEO AFFECTIVE IMPACT PREDICTION WITH MULTIMODAL FUSION AND LONG-SHORT TEMPORAL
 CONTEXT

Yin Zhao,Longjun Cai,Chaoping Tu,Jie Zhang,Wu Wei

Predicting the emotional impact of videos using machine learning is a challengin
g task. Feature extraction, multi-modal fusion and temporal context fusion are c
rucial stages for predicting valence and arousal values in the emotional impact,
 but
have not been successfully exploited. In this paper, we proposed a comprehensive
 framework with innovative designs of model structure and multi-modal fusion str
ategy. We select the most suitable modalities for valence and arousal tasks resp
ectively and each modal feature is extracted using the modality-specific pre-tra
ined deep model on large generic dataset. Two-time-scale structures, one for the
 intra-clip and the other for the inter-clip, are proposed to capture the tempor
al dependency of video content and emotional states. To combine the complementar
y information from multiple modalities, an effective and efficient residual-base
d progressive training strategy is proposed. Each modality is step-wisely combin
ed into the
multi-modal model, responsible for completing the missing parts of features. Wit
h all those above, our proposed prediction framework achieves better performance
 with a large margin compared to the state-of-the-art.
****

Graph inference learning for semi-supervised classification

Chunyan Xu,Zhen Cui,Xiaobin Hong,Tong Zhang,Jian Yang,Wei Liu

In this work, we address the semi-supervised classification of graph data, where
 the categories of those unlabeled nodes are inferred from labeled nodes as well
 as graph structures. Recent works often solve this problem with the advanced gr
aph convolution in a conventional supervised manner, but the performance could b
e heavily affected when labeled data is scarce. Here we propose a Graph Inferenc
e Learning (GIL) framework to boost the performance of node classification by le
arning the inference of node labels on graph topology. To bridge the connection
of two nodes, we formally define a structure relation by encapsulating node attr
ibutes, between-node paths and local topological structures together, which can
make inference conveniently deduced from one node to another node. For learning
the inference process, we further introduce meta-optimization on structure relat
ions from training nodes to validation nodes, such that the learnt graph inferen
ce capability can be better self-adapted into test nodes. Comprehensive evaluati
ons on four benchmark datasets (including Cora, Citeseer, Pubmed and NELL) demon
strate the superiority of our GIL when compared with other state-of-the-art meth
ods in the semi-supervised node classification task.
****

Sparse Coding with Gated Learned ISTA

Kailun Wu,Yiwen Guo,Ziang Li,Changshui Zhang

In this paper, we study the learned iterative shrinkage thresholding algorithm (
LISTA) for solving sparse coding problems. Following assumptions made by prior
works, we first discover that the code components in its estimations may be lowe

r than expected, i.e., require gains, and to address this problem, a gated mecha nism amenable to theoretical analysis is then introduced. Specific design of the gates is inspired by convergence analyses of the mechanism and hence its effect iveness can be formally guaranteed. In addition to the gain gates, we further in troduce overshoot gates for compensating insufficient step size in LISTA. Extens ive empirical results confirm our theoretical findings and verify the effectiven ess of our method.

****

Dimensional Reweighting Graph Convolution Networks

Xu Zou,Qiuye Jia,Jianwei Zhang,Chang Zhou,Zijun Yao,Hongxia Yang,Jie Tang

In this paper, we propose a method named Dimensional reweighting Graph Convoluti onal Networks (DrGCNs), to tackle the problem of variance between dimensional in formation in the node representations of GCNs. We prove that DrGCNs can reduce t he variance of the node representations by connecting our problem to the theory of the mean field. However, practically, we find that the degrees DrGCNs help va ry severely on different datasets. We revisit the problem and develop a new meas ure K to quantify the effect. This measure guides when we should use dimensional reweighting in GCNs and how much it can help. Moreover, it offers insights to e xplain the improvement obtained by the proposed DrGCNs. The dimensional reweight ing block is light-weighted and highly flexible to be built on most of the GCN v ariants. Carefully designed experiments, including several fixes on duplicates, information leaks, and wrong labels of the well-known node classification benchm ark datasets, demonstrate the superior performances of DrGCNs over the existing state-of-the-art approaches. Significant improvements can also be observed on a large scale industrial dataset.

****

ROBUST DISCRIMINATIVE REPRESENTATION LEARNING VIA GRADIENT RESCALING: AN EMPHASI S REGULARISATION PERSPECTIVE

Xinshao Wang,Yang Hua,Elyor Kodirov,Neil M. Robertson

It is fundamental and challenging to train robust and accurate Deep Neural Netwo rks (DNNs) when semantically abnormal examples exist. Although great progress ha s been made, there is still one crucial research question which is not thoroughl y explored yet: What training examples should be focused and how much more shoul d they be emphasised to achieve robust learning? In this work, we study this que stion and propose gradient rescaling (GR) to solve it. GR modifies the magnitude of logit vector’s gradient to emphasise on relatively easier training data poin ts when noise becomes more severe, which functions as explicit emphasis regulari sation to improve the generalisation performance of DNNs. Apart from regularisat ion, we connect GR to examples weighting and designing robust loss functions. We empirically demonstrate that GR is highly anomaly-robust and outperforms the st ate-of-the-art by a large margin, e.g., increasing 7% on CIFAR100 with 40% noisy labels. It is also significantly superior to standard regularisers in both clea n and abnormal settings. Furthermore, we present comprehensive ablation studies to explore the behaviours of GR under different cases, which is informative for applying GR in real-world scenarios.

****

Explaining A Black-box By Using A Deep Variational Information Bottleneck Approa ch

Seojin Bang,Pengtao Xie,Heewook Lee,Wei Wu,Eric Xing

Interpretable machine learning has gained much attention recently. Briefness and comprehensiveness are necessary in order to provide a large amount of informati on concisely when explaining a black-box decision system. However, existing int erpretable machine learning methods fail to consider briefness and comprehensive ness simultaneously, leading to redundant explanations. We propose the variation al information bottleneck for interpretation, VIBI, a system-agnostic interpreta ble method that provides a brief but comprehensive explanation. VIBI adopts an i nformation theoretic principle, information bottleneck principle, as a criterion for finding such explanations. For each instance, VIBI selects key features tha t are maximally compressed about an input (briefness), and informative about a d ecision made by a black-box system on that input (comprehensive). We evaluate VI

BI on three datasets and compare with state-of-the-art interpretable machine lea
rning methods in terms of both interpretability and fidelity evaluated by human
and quantitative metrics.
****

Learning deep graph matching with channel-independent embedding and Hungarian at
tention

Tianshu Yu,Runzhong Wang,Junchi Yan,Baoxin Li

Graph matching aims to establishing node-wise correspondence between two graphs,
 which is a classic combinatorial problem and in general NP-complete. Until very
 recently, deep graph matching methods start to resort to deep networks to achie
ve unprecedented matching accuracy. Along this direction, this paper makes two c
omplementary contributions which can also be reused as plugin in existing works:
 i) a novel node and edge embedding strategy which stimulates the multi-head str
ategy in attention models and allows the information in each channel to be merge
d independently. In contrast, only node embedding is accounted in previous works
; ii) a general masking mechanism over the loss function is devised to improve t
he smoothness of objective learning for graph matching. Using Hungarian algorith
m, it dynamically constructs a structured and sparsely connected layer, taking i
nto account the most contributing matching pairs as hard attention. Our approach
 performs competitively, and can also improve state-of-the-art methods as plugin
, regarding with matching accuracy on three public benchmarks.
****

Out-of-Distribution Detection Using Layerwise Uncertainty in Deep Neural Network
s

Hirono Okamoto,Masahiro Suzuki,Yutaka Matsuo

In this paper, we tackle the problem of detecting samples that are not drawn fro
m the training distribution, i.e., out-of-distribution (OOD) samples, in classif
ication. Many previous studies have attempted to solve this problem by regarding
 samples with low classification confidence as OOD examples using deep neural ne
tworks (DNNs). However, on difficult datasets or models with low classification
ability, these methods incorrectly regard in-distribution samples close to the d
ecision boundary as OOD samples. This problem arises because their approaches us
e only the features close to the output layer and disregard the uncertainty of t
he features. Therefore, we propose a method that extracts the uncertainties of f
eatures in each layer of DNNs using a reparameterization trick and combines them
. In experiments, our method outperforms the existing methods by a large margin,
 achieving state-of-the-art detection performance on several datasets and classi
fication models. For example, our method increases the AUROC score of prior work
 (83.8%) to 99.8% in DenseNet on the CIFAR-100 and Tiny-ImageNet datasets.
****

Semantics Preserving Adversarial Attacks

Ousmane Amadou Dia,Elnaz Barshan,Reza Babanezhad

While progress has been made in crafting visually imperceptible adversarial exam
ples, constructing semantically meaningful ones remains a challenge. In this pap
er, we propose a framework to generate semantics preserving adversarial examples
. First, we present a manifold learning method to capture the semantics of the i
nputs. The motivating principle is to learn the low-dimensional geometric summar
ies of the inputs via statistical inference. Then, we perturb the elements of th
e learned manifold using the Gram-Schmidt process to induce the perturbed elemen
ts to remain in the manifold. To produce adversarial examples, we propose an eff
icient algorithm whereby we leverage the semantics of the inputs as a source of
knowledge upon which we impose adversarial constraints. We apply our approach on
 toy data, images and text, and show its effectiveness in producing semantics pr
eserving adversarial examples which evade existing defenses against adversarial
attacks.
****

Scaling Up Neural Architecture Search with Big Single-Stage Models

Jiahui Yu,Pengchong Jin,Hanxiao Liu,Gabriel Bender,Pieter-Jan Kindermans,Mingxin
g Tan,Thomas Huang,Xiaodan Song,Quoc Le

Neural architecture search (NAS) methods have shown promising results discoverin

g models that are both accurate and fast. For NAS, training a one-shot model has became a popular strategy to approximate the quality of multiple architectures (child models) using a single set of shared weights. To avoid performance degradation due to parameter sharing, most existing methods have a two-stage workflow where the best child model induced from the one-shot model has to be retrained or finetuned. In this work, we propose BigNAS, an approach that simplifies this workflow and scales up neural architecture search to target a wide range of model sizes simultaneously. We propose several techniques to bridge the gap between the distinct initialization and learning dynamics across small and big models with shared parameters, which enable us to train a single-stage model: a single model from which we can directly slice high-quality child models without retraining or finetuning. With BigNAS we are able to train a single set of shared weights on ImageNet and use these weights to obtain child models whose sizes range from 200 to 1000 MFLOPs. Our discovered model family, BigNASModels, achieve top-1 accuracies ranging from 76.5% to 80.9%, surpassing all state-of-the-art models in this range including EfficientNets.

****

AutoSlim: Towards One-Shot Architecture Search for Channel Numbers
Jiahui Yu,Thomas Huang

We study how to set the number of channels in a neural network to achieve better accuracy under constrained resources (e.g., FLOPs, latency, memory footprint or model size). A simple and one-shot approach, named AutoSlim, is presented. Instead of training many network samples and searching with reinforcement learning, we train a single slimmable network to approximate the network accuracy of different channel configurations. We then iteratively evaluate the trained slimmable model and greedily slim the layer with minimal accuracy drop. By this single pass, we can obtain the optimized channel configurations under different resource constraints. We present experiments with MobileNet v1, MobileNet v2, ResNet-50 and RL-searched MNasNet on ImageNet classification. We show significant improvements over their default channel configurations. We also achieve better accuracy than recent channel pruning methods and neural architecture search methods with 100X lower search cost.

Notably, by setting optimized channel numbers, our AutoSlim-MobileNet-v2 at 305M FLOPs achieves 74.2% top-1 accuracy, 2.4% better than default MobileNet-v2 (301M FLOPs), and even 0.2% better than RL-searched MNasNet (317M FLOPs). Our AutoSlim-ResNet-50 at 570M FLOPs, without depthwise convolutions, achieves 1.3% better accuracy than MobileNet-v1 (569M FLOPs).

****

Neural Linear Bandits: Overcoming Catastrophic Forgetting through Likelihood Matching
Tom Zahavy,Shie Mannor

We study neural-linear bandits for solving problems where both exploration and representation learning play an important role. Neural-linear bandits leverage the representation power of deep neural networks and combine it with efficient exploration mechanisms, designed for linear contextual bandits, on top of the last hidden layer. Since the representation is being optimized during learning, information regarding exploration with "old" features is lost. Here, we propose the first limited memory neural-linear bandit that is resilient to this catastrophic forgetting phenomenon. We perform simulations on a variety of real-world problems, including regression, classification, and sentiment analysis, and observe that our algorithm achieves superior performance and shows resilience to catastrophic forgetting.

****

EgoMap: Projective mapping and structured egocentric memory for Deep RL
Edward Beeching,Christian Wolf,Jilles Dibangoye,Olivier Simonin

Tasks involving localization, memorization and planning in partially observable 3D environments are an ongoing challenge in Deep Reinforcement Learning. We pres

ent EgoMap, a spatially structured neural memory architecture. EgoMap augments a deep reinforcement learning agent’s performance in 3D environments on challenging tasks with multi-step objectives. The EgoMap architecture incorporates several inductive biases including a differentiable inverse projection of CNN feature vectors onto a top-down spatially structured map. The map is updated with ego-motion measurements through a differentiable affine transform. We show this architecture outperforms both standard recurrent agents and state of the art agents with structured memory. We demonstrate that incorporating these inductive biases into an agent’s architecture allows for stable training with reward alone, circumventing the expense of acquiring and labelling expert trajectories. A detailed ablation study demonstrates the impact of key aspects of the architecture and through extensive qualitative analysis, we show how the agent exploits its structured internal memory to achieve higher performance.
****

Accelerated Information Gradient flow
Yifei Wang,Wuchen Li
We present a systematic framework for the Nesterov's accelerated gradient flows in the spaces of probabilities embedded with information metrics. Here two metrics are considered, including both the Fisher-Rao metric and the Wasserstein-2 metric. For the Wasserstein-2 metric case, we prove the convergence properties of the accelerated gradient flows, and introduce their formulations in Gaussian families. Furthermore, we propose a practical discrete-time algorithm in particle implementations with an adaptive restart technique. We formulate a novel bandwidth selection method, which learns the Wasserstein-2 gradient direction from Brownian-motion samples. Experimental results including Bayesian inference show the strength of the current method compared with the state-of-the-art.
****

StructPool: Structured Graph Pooling via Conditional Random Fields
Hao Yuan,Shuiwang Ji
Learning high-level representations for graphs is of great importance for graph analysis tasks. In addition to graph convolution, graph pooling is an important but less explored research area. In particular, most of existing graph pooling techniques do not consider the graph structural information explicitly. We argue that such information is important and develop a novel graph pooling technique, know as the StructPool, in this work. We consider the graph pooling as a node clustering problem, which requires the learning of a cluster assignment matrix. We propose to formulate it as a structured prediction problem and employ conditional random fields to capture the relationships among assignments of different nodes. We also generalize our method to incorporate graph topological information in designing the Gibbs energy function. Experimental results on multiple datasets demonstrate the effectiveness of our proposed StructPool.
****

On the Decision Boundaries of Deep Neural Networks: A Tropical Geometry Perspective
Motasem Alfarra,Adel Bibi,Hasan Hammoud,Mohamed Gaafar,Bernard Ghanem
This work tackles the problem of characterizing and understanding the decision boundaries of neural networks with piece-wise linear non-linearity activations. We use tropical geometry, a new development in the area of algebraic geometry, to provide a characterization of the decision boundaries of a simple neural network of the form (Affine, ReLU, Affine). Specifically, we show that the decision boundaries are a subset of a tropical hypersurface, which is intimately related to a polytope formed by the convex hull of two zonotopes. The generators of the zonotopes are precise functions of the neural network parameters. We utilize this geometric characterization to shed light and new perspective on three tasks. In doing so, we propose a new tropical perspective for the lottery ticket hypothesis, where we see the effect of different initializations on the tropical geometric representation of the decision boundaries. Also, we leverage this characterization as a new set of tropical regularizers, which deal directly with the decision boundaries of a network. We investigate the use of these regularizers in neural network pruning (removing network parameters that do not contribute to the

tropical geometric representation of the decision boundaries) and in generating adversarial input attacks (with input perturbations explicitly perturbing the de cision boundaries geometry to change the network prediction of the input).
****

Probabilistic modeling the hidden layers of deep neural networks
Xinjie Lan,Kenneth E. Barner
In this paper, we demonstrate that the parameters of Deep Neural Networks (DNNs) cannot satisfy the i.i.d. prior assumption and activations being i.i.d. is not valid for all the hidden layers of DNNs. Hence, the Gaussian Process cannot corr ectly explain all the hidden layers of DNNs. Alternatively, we introduce a novel probabilistic representation for the hidden layers of DNNs in two aspects: (i) a hidden layer formulates a Gibbs distribution, in which neurons define the ener gy function, and (ii) the connection between two adjacent layers can be modeled by a product of experts model. Based on the probabilistic representation, we dem onstrate that the entire architecture of DNNs can be explained as a Bayesian hie rarchical model. Moreover, the proposed probabilistic representation indicates t hat DNNs have explicit regularizations defined by the hidden layers serving as p rior distributions. Based on the Bayesian explanation for the regularization of DNNs, we propose a novel regularization approach to improve the generalization p erformance of DNNs. Simulation results validate the proposed theories.
****

On the Weaknesses of Reinforcement Learning for Neural Machine Translation
Leshem Choshen,Lior Fox,Zohar Aizenbud,Omri Abend
Reinforcement learning (RL) is frequently used to increase performance in text g eneration tasks,
including machine translation (MT),
notably through the use of Minimum Risk Training (MRT) and Generative Adversaria l Networks (GAN).
However, little is known about what and how these methods learn in the context o f MT.
We prove that one of the most common RL methods for MT does not optimize the expected reward, as well as show that other methods take an infeasibly long time to converge.
In fact, our results suggest that RL practices in MT are likely to improve perfo rmance
only where the pre-trained parameters are already close to yielding the correct translation.
Our findings further suggest that observed gains may be due to effects unrelated to the training signal, concretely, changes in the shape of the distribution cu rve.
****

Stochastically Controlled Compositional Gradient for the Composition problem
Liu Liu,Ji Liu,Cho-Jui Hsieh,Dacheng Tao
We consider composition problems of the form $\frac{1}{n}\sum\nolimits_{i= 1}^ n F_i(\frac{1}{n}\sum\nolimits_{j = 1}^n G_j(x))$. Composition optimization aris es in many important machine learning applications: reinforcement learning, vari ance-aware learning, nonlinear embedding, and many others. Both gradient descent and stochastic gradient descent are straightforward solution, but both require to compute $\frac{1}{n}\sum\nolimits_{j = 1}^n{G_j(x)} $ in each single itera tion, which is inefficient-especially when n is large. Therefore, with the aim of significantly reducing the query complexity of such problems, we designed a stochastically controlled compositional gradient algorithm that incorporates two kinds of variance reduction techniques, and works in both strongly convex and n on-convex settings. The strategy is also accompanied by a mini-batch version of the proposed method that improves query complexity with respect to the size of t he mini-batch. Comprehensive experiments demonstrate the superiority of the prop osed method over existing methods.
****

Sharing Knowledge in Multi-Task Deep Reinforcement Learning
Carlo D'Eramo,Davide Tateo,Andrea Bonarini,Marcello Restelli,Jan Peters

We study the benefit of sharing representations among tasks to enable the effect ive use of deep neural networks in Multi-Task Reinforcement Learning. We leverag e the assumption that learning from different tasks, sharing common properties, is helpful to generalize the knowledge of them resulting in a more effective fea ture extraction compared to learning a single task. Intuitively, the resulting s et of features offers performance benefits when used by Reinforcement Learning a lgorithms. We prove this by providing theoretical guarantees that highlight the conditions for which is convenient to share representations among tasks, extendi ng the well-known finite-time bounds of Approximate Value-Iteration to the multi -task setting. In addition, we complement our analysis by proposing multi-task e xtensions of three Reinforcement Learning algorithms that we empirically evaluat e on widely used Reinforcement Learning benchmarks showing significant improveme nts over the single-task counterparts in terms of sample efficiency and performa nce.

****

HOW IMPORTANT ARE NETWORK WEIGHTS? TO WHAT EXTENT DO THEY NEED AN UPDATE?

Fawaz Sammani,Mahmoud Elsayed,Abdelsalam Hamdi

In the context of optimization, a gradient of a neural network indicates the amo unt a specific weight should change with respect to the loss. Therefore, small g radients indicate a good value of the weight that requires no change and can be kept frozen during training. This paper provides an experimental study on the im portance of a neural network weights, and to which extent do they need to be upd ated. We wish to show that starting from the third epoch, freezing weights which have no informative gradient and are less likely to be changed during training, results in a very slight drop in the overall accuracy (and in sometimes better) . We experiment on the MNIST, CIFAR10 and Flickr8k datasets using several archit ectures (VGG19,
ResNet-110 and DenseNet-121). On CIFAR10, we show that freezing 80% of the VGG19 network parameters from the third epoch onwards results in 0.24% drop in accura cy, while freezing 50% of Resnet-110 parameters results in 0.9% drop in accuracy and finally freezing 70% of Densnet-121 parameters results in 0.57% drop in acc uracy. Furthermore, to experiemnt with real-life applications, we train an image captioning model with attention mechanism on the Flickr8k dataset using LSTM ne tworks, freezing 60% of the parameters from the third epoch onwards, resulting i n a better BLEU-4 score than the fully trained model. Our source code can be fou nd in the appendix.

****

Deep Reasoning Networks: Thinking Fast and Slow, for Pattern De-mixing

Di Chen,Yiwei Bai,Wenting Zhao,Sebastian Ament,John M. Gregoire,Carla P. Gomes

We introduce Deep Reasoning Networks (DRNets), an end-to-end framework that comb ines deep learning with reasoning for solving pattern de-mixing problems, typica lly in an unsupervised or weakly-supervised setting. DRNets exploit problem str ucture and prior knowledge by tightly combining logic and constraint reasoning w ith stochastic-gradient-based neural network optimization. We illustrate the po wer of DRNets on de-mixing overlapping hand-written Sudokus (Multi-MNIST-Sudoku) and on a substantially more complex task in scientific discovery that concerns inferring crystal structures of materials from X-ray diffraction data (Crystal-S tructure-Phase-Mapping). DRNets significantly outperform the state of the art an d experts' capabilities on Crystal-Structure-Phase-Mapping, recovering more prec ise and physically meaningful crystal structures. On Multi-MNIST-Sudoku, DRNets perfectly recovered the mixed Sudokus' digits, with 100% digit accuracy, outperf orming the supervised state-of-the-art MNIST de-mixing models.

****

When Does Self-supervision Improve Few-shot Learning?

Jong-Chyi Su,Subhransu Maji,Bharath Hariharan

We present a technique to improve the generalization of deep representations lea rned on small labeled datasets by introducing self-supervised tasks as auxiliary loss functions. Although recent research has shown benefits of self-supervised learning (SSL) on large unlabeled datasets, its utility on small datasets is unk nown. We find that SSL reduces the relative error rate of few-shot meta-learners

by 4%-27%, even when the datasets are small and only utilizing images within th
e datasets. The improvements are greater when the training set is smaller or the
 task is more challenging. Though the benefits of SSL may increase with larger t
raining sets, we observe that SSL can have a negative impact on performance when
 there is a domain shift between distribution of images used for meta-learning a
nd SSL. Based on this analysis we present a technique that automatically select
images for SSL from a large, generic pool of unlabeled images for a given datase
t using a domain classifier that provides further improvements. We present resul
ts using several meta-learners and self-supervised tasks across datasets with va
rying degrees of domain shifts and label sizes to characterize the effectiveness
 of SSL for few-shot learning.
****

Reinforcement Learning Based Graph-to-Sequence Model for Natural Question Genera
tion

Yu Chen,Lingfei Wu,Mohammed J. Zaki

Natural question generation (QG) aims to generate questions from a passage and a
n answer. Previous works on QG either (i) ignore the rich structure information
hidden in text, (ii) solely rely on cross-entropy loss that leads to issues like
 exposure bias and inconsistency between train/test measurement, or (iii) fail t
o fully exploit the answer information. To address these limitations, in this pa
per, we propose a reinforcement learning (RL) based graph-to-sequence (Graph2Seq
) model for QG. Our model consists of a Graph2Seq generator with a novel Bidirec
tional Gated Graph Neural Network based encoder to embed the passage, and a hybr
id evaluator with a mixed objective combining both cross-entropy and RL losses t
o ensure the generation of syntactically and semantically valid text. We also in
troduce an effective Deep Alignment Network for incorporating the answer informa
tion into the passage at both the word and contextual levels. Our model is end-t
o-end trainable and achieves new state-of-the-art scores, outperforming existing
 methods by a significant margin on the standard SQuAD benchmark.
****

Context-aware Attention Model for Coreference Resolution

Yufei Li,Xiangyu Zhou,Jie Ma,Yu Long,Xuan Wang,Chen Li

Coreference resolution is an important task for gaining more complete understand
ing about texts by artificial intelligence. The state-of-the-art end-to-end neur
al coreference model considers all spans in a document as potential mentions and
 learns to link an antecedent with each possible mention. However, for the verba
tim same mentions, the model tends to get similar or even identical representati
ons based on the features, and this leads to wrongful predictions. In this paper
, we propose to improve the end-to-end system by building an attention model to
reweigh features around different contexts. The proposed model substantially out
performs the state-of-the-art on the English dataset of the CoNLL 2012 Shared Ta
sk with 73.45% F1 score on development data and 72.84% F1 score on test data.
****

SELF: Learning to Filter Noisy Labels with Self-Ensembling

Duc Tam Nguyen,Chaithanya Kumar Mummadi,Thi Phuong Nhung Ngo,Thi Hoai Phuong Ngu
yen,Laura Beggel,Thomas Brox

Deep neural networks (DNNs) have been shown to over-fit a dataset when being tra
ined with noisy labels for a long enough time. To overcome this problem, we pres
ent a simple and effective method self-ensemble label filtering (SELF) to progre
ssively filter out the wrong labels during training. Our method improves the tas
k performance by gradually allowing supervision only from the potentially non-no
isy (clean) labels and stops learning on the filtered noisy labels. For the filt
ering, we form running averages of predictions over the entire training dataset
using the network output at different training epochs. We show that these ensemb
le estimates yield more accurate identification of inconsistent predictions thro
ughout training than the single estimates of the network at the most recent trai
ning epoch. While filtered samples are removed entirely from the supervised trai
ning loss, we dynamically leverage them via semi-supervised learning in the unsu
pervised loss. We demonstrate the positive effect of such an approach on various
 image classification tasks under both symmetric and asymmetric label noise and

at different noise ratios. It substantially outperforms all previous works on noise-aware learning across different datasets and can be applied to a broad set of network architectures.
****

Neural Maximum Common Subgraph Detection with Guided Subgraph Extraction
Yunsheng Bai,Derek Xu,Ken Gu,Xueqing Wu,Agustin Marinovic,Christopher Ro,Yizhou Sun,Wei Wang
Maximum Common Subgraph (MCS) is defined as the largest subgraph that is commonly present in both graphs of a graph pair. Exact MCS detection is NP-hard, and its state-of-the-art exact solver based on heuristic search is slow in practice without any time complexity guarantee. Given the huge importance of this task yet the lack of fast solver, we propose an efficient MCS detection algorithm, Neural MCS, consisting of a novel neural network model that learns the node-node correspondence from the ground-truth MCS result, and a subgraph extraction procedure that uses the neural network output as guidance for final MCS prediction. The whole model guarantees polynomial time complexity with respect to the number of the nodes of the larger of the two input graphs. Experiments on four real graph datasets show that the proposed model is 48.1x faster than the exact solver and more accurate than all the existing competitive approximate approaches to MCS detection.
****

Amharic Negation Handling
Girma Neshir
User generated content contains opinionated texts not only in dominant languages (like English) but also less dominant languages(like Amharic). However, negation handling techniques that supports for sentiment detection is not developed in such less dominant language(i.e. Amharic). Negation handling is one of the challenging tasks for sentiment classification. Thus, this work builds negation handling schemes which enhances Amharic Sentiment classification. The proposed Negation Handling framework combines the lexicon based approach and character ngram based machine learning model. The performance of framework is evaluated using the annotated Amharic News Comments. The system is outperforming the best of all models and the baselines by an accuracy of 98.0. The result is compared with the baselines (without negation handling and word level ngram model).
****

Noise Regularization for Conditional Density Estimation
Jonas Rothfuss,Fabio Ferreira,Simon Boehm,Simon Walther,Maxim Ulrich,Tamim Asfour,Andreas Krause
Modelling statistical relationships beyond the conditional mean is crucial in many settings. Conditional density estimation (CDE) aims to learn the full conditional probability density from data. Though highly expressive, neural network based CDE models can suffer from severe over-fitting when trained with the maximum likelihood objective. Due to the inherent structure of such models, classical regularization approaches in the parameter space are rendered ineffective. To address this issue, we develop a model-agnostic noise regularization method for CDE that adds random perturbations to the data during training. We demonstrate that the proposed approach corresponds to a smoothness regularization and prove its asymptotic consistency. In our experiments, noise regularization significantly and consistently outperforms other regularization methods across seven data sets and three CDE models. The effectiveness of noise regularization makes neural network based CDE the preferable method over previous non- and semi-parametric approaches, even when training data is scarce.
****

Star-Convexity in Non-Negative Matrix Factorization
Johan Bjorck,Carla Gomes,Kilian Weinberger
Non-negative matrix factorization (NMF) is a highly celebrated algorithm for matrix decomposition that guarantees strictly non-negative factors. The underlying optimization problem is computationally intractable, yet in practice gradient descent based solvers often find good solutions. This gap between computational hardness and practical success mirrors recent observations in deep learning, where

it has been the focus of extensive discussion and analysis. In this paper we revisit the NMF optimization problem and analyze its loss landscape in non-worst-case settings. It has recently been observed that gradients in deep networks tend to point towards the final minimizer throughout the optimization. We show that a similar property holds (with high probability) for NMF, provably in a non-worst case model with a planted solution, and empirically across an extensive suite of real-world NMF problems. Our analysis predicts that this property becomes more likely with growing number of parameters, and experiments suggest that a similar trend might also hold for deep neural networks --- turning increasing data sets and models into a blessing from an optimization perspective.
****

Count-guided Weakly Supervised Localization Based on Density Map
Ming Ma,Stephan Chalup,Fayeem Aziz,Yang Liu,Defu Cheng,Zhijian Zhou
Weakly supervised localization (WSL) aims at training a model to find the positions of objects by providing it with only abstract labels. For most of the existing WSL methods, the labels are the class of the main object in an image. In this paper, we generalize WSL to counting machines that apply convolutional neural networks (CNN) and density maps for counting. We show that given only ground-truth count numbers, the density map as a hidden layer can be trained for localizing objects and detecting features. Convolution and pooling are the two major building blocks of CNNs. This paper discusses their impacts on an end-to-end WSL network. The learned features in a density map present in the form of dots. In order to make these features interpretable for human beings, this paper proposes a Gini impurity penalty to regularize the density map. Furthermore, it will be shown that this regularization is similar to the variational term of the β-variational autoencoder. The details of this algorithm are demonstrated through a simple bubble counting task. Finally, the proposed methods are applied to the widely used crowd counting dataset the Mall to learn discriminative features of human figures.
****

Scoring-Aggregating-Planning: Learning task-agnostic priors from interactions and sparse rewards for zero-shot generalization
Huazhe Xu,Boyuan Chen,Yang Gao,Trevor Darrell
Humans can learn task-agnostic priors from interactive experience and utilize the priors for novel tasks without any finetuning. In this paper, we propose Scoring-Aggregating-Planning (SAP), a framework that can learn task-agnostic semantics and dynamics priors from arbitrary quality interactions as well as the corresponding sparse rewards and then plan on unseen tasks in zero-shot condition. The framework finds a neural score function for local regional state and action pairs that can be aggregated to approximate the quality of a full trajectory; moreover, a dynamics model that is learned with self-supervision can be incorporated for planning. Many of previous works that leverage interactive data for policy learning either need massive on-policy environmental interactions or assume access to expert data while we can achieve a similar goal with pure off-policy imperfect data. Instantiating our framework results in a generalizable policy to unseen tasks. Experiments demonstrate that the proposed method can outperform baseline methods on a wide range of applications including gridworld, robotics tasks and video games.
****

SSE-PT: Sequential Recommendation Via Personalized Transformer
Liwei Wu,Shuqing Li,Cho-Jui Hsieh,James Sharpnack
Temporal information is crucial for recommendation problems because user preferences are naturally dynamic in the real world. Recent advances in deep learning, especially the discovery of various attention mechanisms and newer architectures in addition to widely used RNN and CNN in natural language processing, have allowed for better use of the temporal ordering of items that each user has engaged with. In particular, the SASRec model, inspired by the popular Transformer model in natural languages processing, has achieved state-of-the-art results. However, SASRec, just like the original Transformer model, is inherently an un-personalized model and does not include personalized user embeddings. To overcome this

limitation, we propose a Personalized Transformer (SSE-PT) model, outperforming SASRec by almost 5% in terms of NDCG@10 on 5 real-world datasets. Furthermore, after examining some random users' engagement history, we find our model not only more interpretable but also able to focus on recent engagement patterns for each user. Moreover, our SSE-PT model with a slight modification, which we call SSE-PT++, can handle extremely long sequences and outperform SASRec in ranking results with comparable training speed, striking a balance between performance and speed requirements. Our novel application of the Stochastic Shared Embeddings (SSE) regularization is essential to the success of personalization. Code and data are open-sourced at https://github.com/SSE-PT/SSE-PT.
****

Wide Neural Networks are Interpolating Kernel Methods: Impact of Initialization on Generalization
Manuel Nonnenmacher,David Reeb,Ingo Steinwart
The recently developed link between strongly overparametrized neural networks (NNs) and kernel methods has opened a new way to understand puzzling features of NNs, such as their convergence and generalization behaviors. In this paper, we make the bias of initialization on strongly overparametrized NNs under gradient descent explicit. We prove that fully-connected wide ReLU-NNs trained with squared loss are essentially a sum of two parts: The first is the minimum complexity solution of an interpolating kernel method, while the second contributes to the test error only and depends heavily on the initialization. This decomposition has two consequences: (a) the second part becomes negligible in the regime of small initialization variance, which allows us to transfer generalization bounds from minimum complexity interpolating kernel methods to NNs; (b) in the opposite regime, the test error of wide NNs increases significantly with the initialization variance, while still interpolating the training data perfectly. Our work shows that -- contrary to common belief -- the initialization scheme has a strong effect on generalization performance, providing a novel criterion to identify good initialization strategies.
****

Improving Evolutionary Strategies with Generative Neural Networks
Louis Faury,Clément Calauzènes,Olivier Fercoq
Evolutionary Strategies (ES) are a popular family of black-box zeroth-order optimization algorithms which rely on search distributions to efficiently optimize a large variety of objective functions. This paper investigates the potential benefits of using highly flexible search distributions in ES algorithms, in contrast to standard ones (typically Gaussians). We model such distributions with Generative Neural Networks (GNNs) and introduce a new ES algorithm that leverages their expressiveness to accelerate the stochastic search. Because it acts as a plug-in, our approach allows to augment virtually any standard ES algorithm with flexible search distributions. We demonstrate the empirical advantages of this method on a diversity of objective functions.
****

Analysis and Interpretation of Deep CNN Representations as Perceptual Quality Features
Taimoor Tariq,Munchurl Kim
Pre-trained Deep Convolutional Neural Network (CNN) features have popularly been used as full-reference perceptual quality features for CNN based image quality assessment, super-resolution, image restoration and a variety of image-to-image translation problems. In this paper, to get more insight, we link basic human visual perception to characteristics of learned deep CNN representations as a novel and first attempt to interpret them. We characterize the frequency and orientation tuning of channels in trained object detection deep CNNs (e.g., VGG-16) by applying grating stimuli of different spatial frequencies and orientations as input. We observe that the behavior of CNN channels as spatial frequency and orientation selective filters can be used to link basic human visual perception models to their characteristics. Doing so, we develop a theory to get more insight into deep CNN representations as perceptual quality features. We conclude that sensitivity to spatial frequencies that have lower contrast masking thresholds in h

uman visual perception and a definite and strong orientation selectivity are imp
ortant attributes of deep CNN channels that deliver better perceptual quality fe
atures.
****

Program Guided Agent
Shao-Hua Sun,Te-Lin Wu,Joseph J. Lim
Developing agents that can learn to follow natural language instructions has bee
n an emerging research direction. While being accessible and flexible, natural l
anguage instructions can sometimes be ambiguous even to humans. To address this,
 we propose to utilize programs, structured in a formal language, as a precise a
nd expressive way to specify tasks. We then devise a modular framework that lear
ns to perform a task specified by a program – as different circumstances give ri
se to diverse ways to accomplish the task, our framework can perceive which circ
umstance it is currently under, and instruct a multitask policy accordingly to f
ulfill each subtask of the overall task. Experimental results on a 2D Minecraft
environment not only demonstrate that the proposed framework learns to reliably
accomplish program instructions and achieves zero-shot generalization to more co
mplex instructions but also verify the efficiency of the proposed modulation mec
hanism for learning the multitask policy. We also conduct an analysis comparing
various models which learn from programs and natural language instructions in an
 end-to-end fashion.
****

Prestopping: How Does Early Stopping Help Generalization Against Label Noise?
Hwanjun Song,Minseok Kim,Dongmin Park,Jae-Gil Lee
Noisy labels are very common in real-world training data, which lead to poor gen
eralization on test data because of overfitting to the noisy labels. In this pap
er, we claim that such overfitting can be avoided by "early stopping" training a
 deep neural network before the noisy labels are severely memorized. Then, we re
sume training the early stopped network using a "maximal safe set," which mainta
ins a collection of almost certainly true-labeled samples at each epoch since th
e early stop point. Putting them all together, our novel two-phase training meth
od, called Prestopping, realizes noise-free training under any type of label noi
se for practical use. Extensive experiments using four image benchmark data sets
 verify that our method significantly outperforms four state-of-the-art methods
in test error by 0.4–8.2 percent points under existence of real-world noise.
****

Carpe Diem, Seize the Samples Uncertain "at the Moment" for Adaptive Batch Selec
tion
Hwanjun Song,Minseok Kim,Sundong Kim,Jae-Gil Lee
The performance of deep neural networks is significantly affected by how well mi
ni-batches are constructed. In this paper, we propose a novel adaptive batch sel
ection algorithm called Recency Bias that exploits the uncertain samples predict
ed inconsistently in recent iterations. The historical label predictions of each
 sample are used to evaluate its predictive uncertainty within a sliding window.
 By taking advantage of this design, Recency Bias not only accelerates the train
ing step but also achieves a more accurate network. We demonstrate the superiori
ty of Recency Bias by extensive evaluation on two independent tasks. Compared wi
th existing batch selection methods, the results showed that Recency Bias reduce
d the test error by up to 20.5% in a fixed wall-clock training time. At the same
 time, it improved the training time by up to 59.3% to reach the same test error
.
****

Large Batch Optimization for Deep Learning: Training BERT in 76 minutes
Yang You,Jing Li,Sashank Reddi,Jonathan Hseu,Sanjiv Kumar,Srinadh Bhojanapalli,X
iaodan Song,James Demmel,Kurt Keutzer,Cho-Jui Hsieh
Training large deep neural networks on massive datasets is computationally very
 challenging. There has been recent surge in interest in using large batch stoch
astic optimization methods to tackle this issue. The most prominent algorithm in
 this line of research is LARS, which by employing layerwise adaptive learning
rates trains ResNet on ImageNet in a few minutes. However, LARS performs poorly

for attention models like BERT, indicating that its performance gains are not co
nsistent across tasks. In this paper, we first study a principled layerwise adap
tation strategy to accelerate training of deep neural networks using large mini-
batches. Using this strategy, we develop a new layerwise adaptive large batch op
timization technique called LAMB; we then provide convergence analysis of LAMB a
s well as LARS, showing convergence to a stationary point in general nonconvex s
ettings. Our empirical results demonstrate the superior performance of LAMB acro
ss various tasks such as BERT and ResNet-50 training with very little hyperparam
eter tuning. In particular, for BERT training, our optimizer enables use of very
 large batch sizes of 32868 without any degradation of performance. By increasi
ng the batch size to the memory limit of a TPUv3 Pod, BERT training time can be
reduced from 3 days to just 76 minutes.
****