Predicting Floor-Level for 911 Calls with Neural Networks and Smartphone Sensor
Data
William Falcon,Henning Schulzrinne
In cities with tall buildings, emergency responders need an accurate floor level
 location to find 911 callers quickly. We introduce a system to estimate a victi
m's floor level via their mobile device's sensor data in a two-step process. Fir
st, we train a neural network to determine when a smartphone enters or exits a b
uilding via GPS signal changes. Second, we use a barometer equipped smartphone t
o measure the change in barometric pressure from the entrance of the building to
 the victim's indoor location. Unlike impractical previous approaches, our syste
m is the first that does not require the use of beacons, prior knowledge of the
building infrastructure, or knowledge of user behavior. We demonstrate real-worl
d feasibility through 63 experiments across five different tall buildings throug
hout New York City where our system predicted the correct floor level with 100%
accuracy.

****
Some Considerations on Learning to Explore via Meta-Reinforcement Learning
Bradly Stadie,Ge Yang,Rein Houthooft,Xi Chen,Yan Duan,Yuhuai Wu,Pieter Abbeel,Il
ya Sutskever
We consider the problem of exploration in meta reinforcement learning. Two new m
eta reinforcement learning algorithms are suggested: E-MAML and ERL2. Results ar
e presented on a novel environment we call 'Krazy World' and a set of maze envi
ronments. We show E-MAML and ERL2 deliver better performance on tasks where expl
oration is important.
****
MACH: Embarrassingly parallel K-class classification in $O(d\log{K})$ memory a
nd $O(K\log{K} + d\log{K})$ time, instead of $O(Kd)$
Qixuan Huang,Anshumali Shrivastava,Yiqiu Wang
We present Merged-Averaged Classifiers via Hashing (MACH) for K-classification
 with large K. Compared to traditional one-vs-all classifiers that require $O(
Kd)$ memory and inference cost, MACH only need $O(d\log{K})$ memory while only r
equiring $O(K\log{K} + d\log{K})$ operation for inference. MACH is the first gen
eric K-classification algorithm, with provably theoretical guarantees, which r
equires $O(\log{K})$ memory without any assumption on the relationship between c
lasses. MACH uses universal hashing to reduce classification with a large number
 of classes to few independent classification task with very small (constant) nu
mber of classes. We provide theoretical quantification of accuracy-memory tradeo
ff by showing the first connection between extreme classification and heavy hitt
ers. With MACH we can train ODP dataset with 100,000 classes and 400,000 feature
s on a single Titan X GPU (12GB), with the classification accuracy of 19.28\%, w
hich is the best-reported accuracy on this dataset. Before this work, the best p
erforming baseline is a one-vs-all classifier that requires 40 billion parameter
s (320 GB model size) and achieves 9\% accuracy. In contrast, MACH can achieve
9\% accuracy with 480x reduction in the model size (of mere 0.6GB). With MACH, w
e also demonstrate complete training of fine-grained imagenet dataset (compresse
d size 104GB), with 21,000 classes, on a single GPU.
****
Deterministic Policy Imitation Gradient Algorithm
Fumihiro Sasaki,Atsuo Kawaguchi
The goal of imitation learning (IL) is to enable a learner to imitate an expert’
s behavior given the expert’s demonstrations. Recently, generative adversarial i
mitation learning (GAIL) has successfully achieved it even on complex continuous
 control tasks. However, GAIL requires a huge number of interactions with enviro
nment during training. We believe that IL algorithm could be more applicable to
the real-world environments if the number of interactions could be reduced. To t
his end, we propose a model free, off-policy IL algorithm for continuous control
. The keys of our algorithm are two folds: 1) adopting deterministic policy that
 allows us to derive a novel type of policy gradient which we call deterministic
 policy imitation gradient (DPIG), 2) introducing a function which we call state

screening function (SSF) to avoid noisy policy updates with states that are not typical of those appeared on the expert's demonstrations. Experimental results show that our algorithm can achieve the goal of IL with at least tens of times less interactions than GAIL on a variety of continuous control tasks.
****

Searching for Activation Functions
Prajit Ramachandran,Barret Zoph,Quoc V. Le
The choice of activation functions in deep networks has a significant effect on the training dynamics and task performance. Currently, the most successful and widely-used activation function is the Rectified Linear Unit (ReLU). Although various hand-designed alternatives to ReLU have been proposed, none have managed to replace it due to inconsistent gains. In this work, we propose to leverage automatic search techniques to discover new activation functions. Using a combination of exhaustive and reinforcement learning-based search, we discover multiple novel activation functions. We verify the effectiveness of the searches by conducting an empirical evaluation with the best discovered activation function. Our experiments show that the best discovered activation function, f(x) = x * sigmoid(beta * x), which we name Swish, tends to work better than ReLU on deeper models across a number of challenging datasets. For example, simply replacing ReLUs with Swish units improves top-1 classification accuracy on ImageNet by 0.9% for Mobile NASNet-A and 0.6% for Inception-ResNet-v2. The simplicity of Swish and its similarity to ReLU make it easy for practitioners to replace ReLUs with Swish units in any neural network.
****

Improving Search Through A3C Reinforcement Learning Based Conversational Agent
Milan Aggarwal,Aarushi Arora,Shagun Sodhani,Balaji Krishnamurthy
We develop a reinforcement learning based search assistant which can assist users through a set of actions and sequence of interactions to enable them realize their intent. Our approach caters to subjective search where the user is seeking digital assets such as images which is fundamentally different from the tasks which have objective and limited search modalities. Labeled conversational data is generally not available in such search tasks and training the agent through human interactions can be time consuming. We propose a stochastic virtual user which impersonates a real user and can be used to sample user behavior efficiently to train the agent which accelerates the bootstrapping of the agent. We develop A3C algorithm based context preserving architecture which enables the agent to provide contextual assistance to the user. We compare the A3C agent with Q-learning and evaluate its performance on average rewards and state values it obtains with the virtual user in validation episodes. Our experiments show that the agent learns to achieve higher rewards and better states.
****

Identifying Analogies Across Domains
Yedid Hoshen,Lior Wolf
Identifying analogies across domains without supervision is a key task for artificial intelligence. Recent advances in cross domain image mapping have concentrated on translating images across domains. Although the progress made is impressive, the visual fidelity many times does not suffice for identifying the matching sample from the other domain. In this paper, we tackle this very task of finding exact analogies between datasets i.e. for every image from domain A find an analogous image in domain B. We present a matching-by-synthesis approach: AN-GAN, and show that it outperforms current techniques. We further show that the cross-domain mapping task can be broken into two parts: domain alignment and learning the mapping function. The tasks can be iteratively solved, and as the alignment is improved, the unsupervised translation function reaches quality comparable to full supervision.
****

Bi-Directional Block Self-Attention for Fast and Memory-Efficient Sequence Modeling
Tao Shen,Tianyi Zhou,Guodong Long,Jing Jiang,Chengqi Zhang
Recurrent neural networks (RNN), convolutional neural networks (CNN) and self-at

tention networks (SAN) are commonly used to produce context-aware representation s. RNN can capture long-range dependency but is hard to parallelize and not time -efficient. CNN focuses on local dependency but does not perform well on some ta sks. SAN can model both such dependencies via highly parallelizable computation, but memory requirement grows rapidly in line with sequence length. In this pape r, we propose a model, called "bi-directional block self-attention network (Bi-B loSAN)", for RNN/CNN-free sequence encoding. It requires as little memory as RNN but with all the merits of SAN. Bi-BloSAN splits the entire sequence into block s, and applies an intra-block SAN to each block for modeling local context, then applies an inter-block SAN to the outputs for all blocks to capture long-range dependency. Thus, each SAN only needs to process a short sequence, and only a sm all amount of memory is required. Additionally, we use feature-level attention t o handle the variation of contexts around the same word, and use forward/backwar d masks to encode temporal order information. On nine benchmark datasets for dif ferent NLP tasks, Bi-BloSAN achieves or improves upon state-of-the-art accuracy, and shows better efficiency-memory trade-off than existing RNN/CNN/SAN.
****

WHAI: Weibull Hybrid Autoencoding Inference for Deep Topic Modeling
Hao Zhang,Bo Chen,Dandan Guo,Mingyuan Zhou
To train an inference network jointly with a deep generative topic model, making it both scalable to big corpora and fast in out-of-sample prediction, we develo p Weibull hybrid autoencoding inference (WHAI) for deep latent Dirichlet allocat ion, which infers posterior samples via a hybrid of stochastic-gradient MCMC and autoencoding variational Bayes. The generative network of WHAI has a hierarchy of gamma distributions, while the inference network of WHAI is a Weibull upward-downward variational autoencoder, which integrates a deterministic-upward deep n eural network, and a stochastic-downward deep generative model based on a hierar chy of Weibull distributions. The Weibull distribution can be used to well appro ximate a gamma distribution with an analytic Kullback-Leibler divergence, and ha s a simple reparameterization via the uniform noise, which help efficiently comp ute the gradients of the evidence lower bound with respect to the parameters of the inference network. The effectiveness and efficiency of WHAI are illustrated with experiments on big corpora.

****

The loss surface and expressivity of deep convolutional neural networks
Quynh Nguyen,Matthias Hein
We analyze the expressiveness and loss surface of practical deep convolutional
neural networks (CNNs) with shared weights and max pooling layers. We show
that such CNNs produce linearly independent features at a “wide” layer which
has more neurons than the number of training samples. This condition holds e.g.
for the VGG network. Furthermore, we provide for such wide CNNs necessary
and sufficient conditions for global minima with zero training error. For the ca se
where the wide layer is followed by a fully connected layer we show that almost
every critical point of the empirical loss is a global minimum with zero trainin g
error. Our analysis suggests that both depth and width are very important in dee p
learning. While depth brings more representational power and allows the network
to learn high level features, width smoothes the optimization landscape of the
loss function in the sense that a sufficiently wide network has a well-behaved l oss
surface with almost no bad local minima.
****

Seq2SQL: Generating Structured Queries From Natural Language Using Reinforcement Learning
Victor Zhong,Caiming Xiong,Richard Socher
Relational databases store a significant amount of the worlds data. However, acc essing this data currently requires users to understand a query language such as

SQL. We propose Seq2SQL, a deep neural network for translating natural language questions to corresponding SQL queries. Our model uses rewards from in the loop query execution over the database to learn a policy to generate the query, which contains unordered parts that are less suitable for optimization via cross entropy loss. Moreover, Seq2SQL leverages the structure of SQL to prune the space of generated queries and significantly simplify the generation problem. In addition to the model, we release WikiSQL, a dataset of 80654 hand-annotated examples of questions and SQL queries distributed across 24241 tables fromWikipedia that is an order of magnitude larger than comparable datasets. By applying policy based reinforcement learning with a query execution environment to WikiSQL, Seq2SQL outperforms a state-of-the-art semantic parser, improving execution accuracy from 35.9% to 59.4% and logical form accuracy from 23.4% to 48.3%.
****
Recursive Binary Neural Network Learning Model with 2-bit/weight Storage Requirement
Tianchan Guan,Xiaoyang Zeng,Mingoo Seok
This paper presents a storage-efficient learning model titled Recursive Binary Neural Networks for embedded and mobile devices having a limited amount of on-chip data storage such as hundreds of kilo-Bytes. The main idea of the proposed model is to recursively recycle data storage of weights (parameters) during training. This enables a device with a given storage constraint to train and instantiate a neural network classifier with a larger number of weights on a chip, achieving better classification accuracy. Such efficient use of on-chip storage reduces off-chip storage accesses, improving energy-efficiency and speed of training. We verified the proposed training model with deep and convolutional neural network classifiers on the MNIST and voice activity detection benchmarks. For the deep neural network, our model achieves data storage requirement of as low as 2 bits/weight, whereas the conventional binary neural network learning models require data storage of 8 to 32 bits/weight. With the same amount of data storage, our model can train a bigger network having more weights, achieving 1% less test error than the conventional binary neural network learning model. To achieve the similar classification error, the conventional binary neural network model requires 4× more data storage for weights than our proposed model. For the convolution neural network classifier, the proposed model achieves 2.4% less test error for the same on-chip storage or 6× storage savings to achieve the similar accuracy.

****
Learning to select examples for program synthesis
Yewen Pu,Zachery Miranda,Armando Solar-Lezama,Leslie Pack Kaelbling
Program synthesis is a class of regression problems where one seeks a solution, in the form of a source-code program, that maps the inputs to their corresponding outputs exactly. Due to its precise and combinatorial nature, it is commonly formulated as a constraint satisfaction problem, where input-output examples are expressed constraints, and solved with a constraint solver. A key challenge of this formulation is that of scalability: While constraint solvers work well with few well-chosen examples, constraining the entire set of example constitutes a significant overhead in both time and memory. In this paper we address this challenge by constructing a representative subset of examples that is both small and is able to constrain the solver sufficiently. We build the subset one example at a time, using a trained discriminator to predict the probability of unchosen input-output examples conditioned on the chosen input-output examples, adding the least probable example to the subset. Experiment on a diagram drawing domain shows our approach produces subset of examples that are small and representative for the constraint solver.
****
Adversarial Learning for Semi-Supervised Semantic Segmentation
Wei-Chih Hung,Yi-Hsuan Tsai,Yan-Ting Liou,Yen-Yu Lin,Ming-Hsuan Yang
We propose a method for semi-supervised semantic segmentation using the adversarial network. While most existing discriminators are trained to classify input images as real or fake on the image level, we design a discriminator in a fully co

nvolutional manner to differentiate the predicted probability maps from the ground truth segmentation distribution with the consideration of the spatial resolution. We show that the proposed discriminator can be used to improve the performance on semantic segmentation by coupling the adversarial loss with the standard cross entropy loss on the segmentation network. In addition, the fully convolutional discriminator enables the semi-supervised learning through discovering the trustworthy regions in prediction results of unlabeled images, providing additional supervisory signals. In contrast to existing methods that utilize weakly-labeled images, our method leverages unlabeled images without any annotation to enhance the segmentation model. Experimental results on both the PASCAL VOC 2012 dataset and the Cityscapes dataset demonstrate the effectiveness of our algorithm.
****

The Information-Autoencoding Family: A Lagrangian Perspective on Latent Variable Generative Modeling

Shengjia Zhao,Jiaming Song,Stefano Ermon

A variety of learning objectives have been recently proposed for training generative models. We show that many of them, including InfoGAN, ALI/BiGAN, ALICE, CycleGAN, VAE, β-VAE, adversarial autoencoders, AVB, and InfoVAE, are Lagrangian duals of the same primal optimization problem. This generalization reveals the implicit modeling trade-offs between flexibility and computational requirements being made by these models. Furthermore, we characterize the class of all objectives that can be optimized under certain computational constraints.
Finally, we show how this new Lagrangian perspective can explain undesirable behavior of existing methods and provide new principled solutions.
****

Defense-GAN: Protecting Classifiers Against Adversarial Attacks Using Generative Models

Pouya Samangouei,Maya Kabkab,Rama Chellappa

In recent years, deep neural network approaches have been widely adopted for machine learning tasks, including classification. However, they were shown to be vulnerable to adversarial perturbations: carefully crafted small perturbations can cause misclassification of legitimate images. We propose Defense-GAN, a new framework leveraging the expressive capability of generative models to defend deep neural networks against such attacks. Defense-GAN is trained to model the distribution of unperturbed images. At inference time, it finds a close output to a given image which does not contain the adversarial changes. This output is then fed to the classifier. Our proposed method can be used with any classification model and does not modify the classifier structure or training procedure. It can also be used as a defense against any attack as it does not assume knowledge of the process for generating the adversarial examples. We empirically show that Defense-GAN is consistently effective against different attack methods and improves on existing defense strategies.
****

Ground-Truth Adversarial Examples

Nicholas Carlini,Guy Katz,Clark Barrett,David L. Dill

The ability to deploy neural networks in real-world, safety-critical systems is severely limited by the presence of adversarial examples: slightly perturbed inputs that are misclassified by the network. In recent years, several techniques have been proposed for training networks that are robust to such examples; and each time stronger attacks have been devised, demonstrating the shortcomings of existing defenses. This highlights a key difficulty in designing an effective defense: the inability to assess a network's robustness against future attacks. We propose to address this difficulty through formal verification techniques. We construct ground truths: adversarial examples with a provably-minimal distance from a given input point. We demonstrate how ground truths can serve to assess the effectiveness of attack techniques, by comparing the adversarial examples produced by those attacks to the ground truths; and also of defense techniques, by computing the distance to the ground truths before and after the defense is applied, and measuring the improvement. We use this technique to assess recently suggested attack and defense techniques.

****
CNNs as Inverse Problem Solvers and Double Network Superresolution
Cem TARHAN,Gözde BOZDA■I AKAR
In recent years Convolutional Neural Networks (CNN) have been used extensively for Superresolution (SR). In this paper, we use inverse problem and sparse representation solutions to form a mathematical basis for CNN operations. We show how a single neuron is able to provide the optimum solution for inverse problem, given a low resolution image dictionary as an operator. Introducing a new concept called Representation Dictionary Duality, we show that CNN elements (filters) are trained to be representation vectors and then, during reconstruction, used as dictionaries. In the light of theoretical work, we propose a new algorithm which uses two networks with different structures that are separately trained with low and high coherency image patches and show that it performs faster compared to the state-of-the-art algorithms while not sacrificing from performance.
****
Unsupervised Learning of Goal Spaces for Intrinsically Motivated Goal Exploration
Alexandre Péré,Sébastien Forestier,Olivier Sigaud,Pierre-Yves Oudeyer
Intrinsically motivated goal exploration algorithms enable machines to discover repertoires of policies that produce a diversity of effects in complex environments. These exploration algorithms have been shown to allow real world robots to acquire skills such as tool use in high-dimensional continuous state and action spaces. However, they have so far assumed that self-generated goals are sampled in a specifically engineered feature space, limiting their autonomy. In this work, we propose an approach using deep representation learning algorithms to learn an adequate goal space. This is a developmental 2-stage approach: first, in a perceptual learning stage, deep learning algorithms use passive raw sensor observations of world changes to learn a corresponding latent space; then goal exploration happens in a second stage by sampling goals in this latent space. We present experiments with a simulated robot arm interacting with an object, and we show that exploration algorithms using such learned representations can closely match, and even sometimes improve, the performance obtained using engineered representations.
****
Generation and Consolidation of Recollections for Efficient Deep Lifelong Learning
Matt Riemer,Michele Franceschini,and Tim Klinger
Deep lifelong learning systems need to efficiently manage resources to scale to large numbers of experiences and non-stationary goals. In this paper, we explore the relationship between lossy compression and the resource constrained lifelong learning problem of function transferability. We demonstrate that lossy episodic experience storage can enable efficient function transferability between different architectures and algorithms at a fraction of the storage cost of lossless storage. This is achieved by introducing a generative knowledge distillation strategy that does not store any full training examples. As an important extension of this idea, we show that lossy recollections stabilize deep networks much better than lossless sampling in resource constrained settings of lifelong learning while avoiding catastrophic forgetting. For this setting, we propose a novel dual purpose recollection buffer used to both stabilize the recollection generator itself and an accompanying reasoning model.
****
Word translation without parallel data
Guillaume Lample,Alexis Conneau,Marc'Aurelio Ranzato,Ludovic Denoyer,Hervé Jégou
State-of-the-art methods for learning cross-lingual word embeddings have relied on bilingual dictionaries or parallel corpora. Recent studies showed that the need for parallel data supervision can be alleviated with character-level information. While these methods showed encouraging results, they are not on par with their supervised counterparts and are limited to pairs of languages sharing a common alphabet. In this work, we show that we can build a bilingual dictionary betw

een two languages without using any parallel corpora, by aligning monolingual word embedding spaces in an unsupervised way. Without using any character information, our model even outperforms existing supervised methods on cross-lingual tasks for some language pairs. Our experiments demonstrate that our method works very well also for distant language pairs, like English-Russian or English-Chinese. We finally describe experiments on the English-Esperanto low-resource language pair, on which there only exists a limited amount of parallel data, to show the potential impact of our method in fully unsupervised machine translation. Our code, embeddings and dictionaries are publicly available.

****

Towards Provable Control for Unknown Linear Dynamical Systems

Sanjeev Arora,Elad Hazan,Holden Lee,Karan Singh,Cyril Zhang,Yi Zhang

We study the control of symmetric linear dynamical systems with unknown dynamics and a hidden state. Using a recent spectral filtering technique for concisely representing such systems in a linear basis, we formulate optimal control in this setting as a convex program. This approach eliminates the need to solve the non-convex problem of explicit identification of the system and its latent state, and allows for provable optimality guarantees for the control signal. We give the first efficient algorithm for finding the optimal control signal with an arbitrary time horizon T, with sample complexity (number of training rollouts) polynomial only in log(T) and other relevant parameters.

****

Critical Points of Linear Neural Networks: Analytical Forms and Landscape Properties

Yi Zhou,Yingbin Liang

Due to the success of deep learning to solving a variety of challenging machine learning tasks, there is a rising interest in understanding loss functions for training neural networks from a theoretical aspect. Particularly, the properties of critical points and the landscape around them are of importance to determine the convergence performance of optimization algorithms. In this paper, we provide a necessary and sufficient characterization of the analytical forms for the critical points (as well as global minimizers) of the square loss functions for linear neural networks. We show that the analytical forms of the critical points characterize the values of the corresponding loss functions as well as the necessary and sufficient conditions to achieve global minimum. Furthermore, we exploit the analytical forms of the critical points to characterize the landscape properties for the loss functions of linear neural networks and shallow ReLU networks. One particular conclusion is that: While the loss function of linear networks has no spurious local minimum, the loss function of one-hidden-layer nonlinear networks with ReLU activation function does have local minimum that is not global minimum.

****

WSNet: Learning Compact and Efficient Networks with Weight Sampling

Xiaojie Jin,Yingzhen Yang,Ning Xu,Jianchao Yang,Jiashi Feng,Shuicheng Yan

■We present a new approach and a novel architecture, termed WSNet, for learning compact and efficient deep neural networks. Existing approaches conventionally learn full model parameters independently and then compress them via \emph{ad hoc} processing such as model pruning or filter factorization. Alternatively, WSNet proposes learning model parameters by sampling from a compact set of learnable parameters, which naturally enforces {parameter sharing} throughout the learning process. We demonstrate that such a novel weight sampling approach (and induced WSNet) promotes both weights and computation sharing favorably. By employing this method, we can more efficiently learn much smaller networks with competitive performance compared to baseline networks with equal numbers of convolution filters. Specifically, we consider learning compact and efficient 1D convolutional neural networks for audio classification. Extensive experiments on multiple audio classification datasets verify the effectiveness of WSNet. Combined with weight quantization, the resulted models are up to \textbf{180$\times$} smaller and theoretically up to \textbf{16$\times$} faster than the well-established baselines, without noticeable performance drop.

****
Meta-Learning and Universality: Deep Representations and Gradient Descent can Approximate any Learning Algorithm
Chelsea Finn,Sergey Levine
Learning to learn is a powerful paradigm for enabling models to learn from data more effectively and efficiently. A popular approach to meta-learning is to train a recurrent model to read in a training dataset as input and output the parameters of a learned model, or output predictions for new test inputs. Alternatively, a more recent approach to meta-learning aims to acquire deep representations that can be effectively fine-tuned, via standard gradient descent, to new tasks. In this paper, we consider the meta-learning problem from the perspective of universality, formalizing the notion of learning algorithm approximation and comparing the expressive power of the aforementioned recurrent models to the more recent approaches that embed gradient descent into the meta-learner. In particular, we seek to answer the following question: does deep representation combined with standard gradient descent have sufficient capacity to approximate any learning algorithm? We find that this is indeed true, and further find, in our experiments, that gradient-based meta-learning consistently leads to learning strategies that generalize more widely compared to those represented by recurrent models.
****
Maximum a Posteriori Policy Optimisation
Abbas Abdolmaleki,Jost Tobias Springenberg,Yuval Tassa,Remi Munos,Nicolas Heess,Martin Riedmiller
We introduce a new algorithm for reinforcement learning called Maximum a-posteriori Policy Optimisation (MPO) based on coordinate ascent on a relative-entropy objective. We show that several existing methods can directly be related to our derivation. We develop two off-policy algorithms and demonstrate that they are competitive with the state-of-the-art in deep reinforcement learning. In particular, for continuous control, our method outperforms existing methods with respect to sample efficiency, premature convergence and robustness to hyperparameter settings.
****
Trace norm regularization and faster inference for embedded speech recognition RNNs
Markus Kliegl,Siddharth Goyal,Kexin Zhao,Kavya Srinet,Mohammad Shoeybi
We propose and evaluate new techniques for compressing and speeding up dense matrix multiplications as found in the fully connected and recurrent layers of neural networks for embedded large vocabulary continuous speech recognition (LVCSR). For compression, we introduce and study a trace norm regularization technique for training low rank factored versions of matrix multiplications. Compared to standard low rank training, we show that our method leads to good accuracy versus number of parameter trade-offs and can be used to speed up training of large models. For speedup, we enable faster inference on ARM processors through new open sourced kernels optimized for small batch sizes, resulting in 3x to 7x speed ups over the widely used gemmlowp library. Beyond LVCSR, we expect our techniques and kernels to be more generally applicable to embedded neural networks with large fully connected or recurrent layers.
****
Domain Adaptation for Deep Reinforcement Learning in Visually Distinct Games
Dino S. Ratcliffe,Luca Citi,Sam Devlin,Udo Kruschwitz
Many deep reinforcement learning approaches use graphical state representations, this means visually distinct games that share the same underlying structure cannot
effectively share knowledge. This paper outlines a new approach for learning underlying game state embeddings irrespective of the visual rendering of the game
state. We utilise approaches from multi-task learning and domain adaption in order to place visually distinct game states on a shared embedding manifold. We present our results in the context of deep reinforcement learning agents.
****

The Implicit Bias of Gradient Descent on Separable Data
Daniel Soudry,Elad Hoffer,Mor Shpigel Nacson,Nathan Srebro
We show that gradient descent on an unregularized logistic regression
problem, for almost all separable datasets, converges to the same direction as t
he max-margin solution. The result generalizes also to other monotone decreasing
 loss functions with an infimum at infinity, and we also discuss a multi-class g
eneralizations to the cross entropy loss. Furthermore,
we show this convergence is very slow, and only logarithmic in the
convergence of the loss itself. This can help explain the benefit
of continuing to optimize the logistic or cross-entropy loss even
after the training error is zero and the training loss is extremely
small, and, as we show, even if the validation loss increases. Our
methodology can also aid in understanding implicit regularization
in more complex models and with other optimization methods.
****
Online Learning Rate Adaptation with Hypergradient Descent
Atilim Gunes Baydin,Robert Cornish,David Martinez Rubio,Mark Schmidt,Frank Wood
We introduce a general method for improving the convergence rate of gradient-bas
ed optimizers that is easy to implement and works well in practice. We demonstr
ate the effectiveness of the method in a range of optimization problems by apply
ing it to stochastic gradient descent, stochastic gradient descent with Nesterov
 momentum, and Adam, showing that it significantly reduces the need for the manu
al tuning of the initial learning rate for these commonly used algorithms. Our
method works by dynamically updating the learning rate during optimization using
 the gradient with respect to the learning rate of the update rule itself. Comp
uting this "hypergradient" needs little additional computation, requires only on
e extra copy of the original gradient to be stored in memory, and relies upon no
thing more than what is provided by reverse-mode automatic differentiation.
****
Learning Parsimonious Deep Feed-forward Networks
Zhourong Chen,Xiaopeng Li,Nevin L. Zhang
Convolutional neural networks and recurrent neural networks are designed with ne
twork structures well suited to the nature of spacial and sequential data respec
tively. However, the structure of standard feed-forward neural networks (FNNs) i
s simply a stack of fully connected layers, regardless of the feature correlatio
ns in data. In addition, the number of layers and the number of neurons are manu
ally tuned on validation data, which is time-consuming and may lead to suboptima
l networks. In this paper, we propose an unsupervised structure learning method
for learning parsimonious deep FNNs. Our method determines the number of layers,
 the number of neurons at each layer, and the sparse connectivity between adjace
nt layers automatically from data. The resulting models are called Backbone-Skip
path Neural Networks (BSNNs). Experiments on 17 tasks show that, in comparison w
ith FNNs, BSNNs can achieve better or comparable classification performance wit
h much fewer parameters. The interpretability of BSNNs is also shown to be bette
r than that of FNNs.
****
Latent forward model for Real-time Strategy game planning with incomplete inform
ation
Yuandong Tian,Qucheng Gong
Model-free deep reinforcement learning approaches have shown superhuman performa
nce in simulated environments (e.g., Atari games, Go, etc). During training, the
se approaches often implicitly construct a latent space that contains key inform
ation for decision making. In this paper, we learn a forward model on this laten
t space and apply it to model-based planning in miniature Real-time Strategy gam
e with incomplete information (MiniRTS). We first show that the latent space con
structed from existing actor-critic models contains relevant information of the
game, and design training procedure to learn forward models. We also show that o
ur learned forward model can predict meaningful future state and is usable for l
atent space Monte-Carlo Tree Search (MCTS), in terms of win rates against rule-b
ased agents.

****
Learning Weighted Representations for Generalization Across Designs
Fredrik D. Johansson,Nathan Kallus,Uri Shalit,David Sontag
Predictive models that generalize well under distributional shift are often desi
rable and sometimes crucial to machine learning applications. One example is the
 estimation of treatment effects from observational data, where a subtask is to
predict the effect of a treatment on subjects that are systematically different
from those who received the treatment in the data. A related kind of distributio
nal shift appears in unsupervised domain adaptation, where we are tasked with ge
neralizing to a distribution of inputs that is different from the one in which w
e observe labels. We pose both of these problems as prediction under a shift in
design. Popular methods for overcoming distributional shift are often heuristic
or rely on assumptions that are rarely true in practice, such as having a well-s
pecified model or knowing the policy that gave rise to the observed data. Other
methods are hindered by their need for a pre-specified metric for comparing obse
rvations, or by poor asymptotic properties. In this work, we devise a bound on t
he generalization error under design shift, based on integral probability metric
s and sample re-weighting. We combine this idea with representation learning, ge
neralizing and tightening existing results in this space. Finally, we propose an
 algorithmic framework inspired by our bound and verify is effectiveness in caus
al effect estimation.
****
Spherical CNNs
Taco S. Cohen,Mario Geiger,Jonas Köhler,Max Welling
Convolutional Neural Networks (CNNs) have become the method of choice for learni
ng problems involving 2D planar images. However, a number of problems of recent
interest have created a demand for models that can analyze spherical images. Exa
mples include omnidirectional vision for drones, robots, and autonomous cars, mo
lecular regression problems, and global weather and climate modelling. A naive a
pplication of convolutional networks to a planar projection of the spherical sig
nal is destined to fail, because the space-varying distortions introduced by suc
h a projection will make translational weight sharing ineffective.

In this paper we introduce the building blocks for constructing spherical CNNs.
We propose a definition for the spherical cross-correlation that is both express
ive and rotation-equivariant. The spherical correlation satisfies a generalized
Fourier theorem, which allows us to compute it efficiently using a generalized (
non-commutative) Fast Fourier Transform (FFT) algorithm. We demonstrate the comp
utational efficiency, numerical accuracy, and effectiveness of spherical CNNs ap
plied to 3D model recognition and atomization energy regression.
****
Syntax-Directed Variational Autoencoder for Structured Data
Hanjun Dai,Yingtao Tian,Bo Dai,Steven Skiena,Le Song
Deep generative models have been enjoying success in modeling continuous data. H
owever it remains challenging to capture the representations for discrete struct
ures with formal grammars and semantics, e.g., computer programs and molecular s
tructures. How to generate both syntactically and semantically correct data stil
l remains largely an open problem. Inspired by the theory of compiler where synt
ax and semantics check is done via syntax-directed translation (SDT), we propose
 a novel syntax-directed variational autoencoder (SD-VAE) by introducing stochas
tic lazy attributes. This approach converts the offline SDT check into on-the-fl
y generated guidance for constraining the decoder. Comparing to the state-of-the
-art methods, our approach enforces constraints on the output space so that the
output will be not only syntactically valid, but also semantically reasonable. W
e evaluate the proposed model with applications in programming language and mole
cules, including reconstruction and program/molecule optimization. The results d
emonstrate the effectiveness in incorporating syntactic and semantic constraints
 in discrete generative models, which is significantly better than current state
-of-the-art approaches.
****

HybridNet: A Hybrid Neural Architecture to Speed-up Autoregressive Models
Yanqi Zhou,Wei Ping,Sercan Arik,Kainan Peng,Greg Diamos
This paper introduces HybridNet, a hybrid neural network to speed-up autoregress ive
models for raw audio waveform generation. As an example, we propose
a hybrid model that combines an autoregressive network named WaveNet and a
conventional LSTM model to address speech synthesis. Instead of generating
one sample per time-step, the proposed HybridNet generates multiple samples per
time-step by exploiting the long-term memory utilization property of LSTMs. In
the evaluation, when applied to text-to-speech, HybridNet yields state-of-art pe rformance.
HybridNet achieves a 3.83 subjective 5-scale mean opinion score on
US English, largely outperforming the same size WaveNet in terms of naturalness
and provide 2x speed up at inference.
****
Learning to Count Objects in Natural Images for Visual Question Answering
Yan Zhang,Jonathon Hare,Adam Prügel-Bennett
Visual Question Answering (VQA) models have struggled with counting objects in n atural images so far. We identify a fundamental problem due to soft attention in these models as a cause. To circumvent this problem, we propose a neural networ k component that allows robust counting from object proposals. Experiments on a toy task show the effectiveness of this component and we obtain state-of-the-art accuracy on the number category of the VQA v2 dataset without negatively affect ing other categories, even outperforming ensemble models with our single model. On a difficult balanced pair metric, the component gives a substantial improveme nt in counting over a strong baseline by 6.6%.
****
Key Protected Classification for GAN Attack Resilient Collaborative Learning
Mert Bülent Sar■y■ld■z,Ramazan Gökberk Cinbi■,Erman Ayday
Large-scale publicly available datasets play a fundamental role in training deep learning models. However, large-scale
datasets are difficult to collect in problems that involve processing of sensiti ve information.
Collaborative learning techniques provide a privacy-preserving solution in such cases, by enabling
training over a number of private datasets that are not shared by their owners. Existing collaborative learning
techniques, combined with differential privacy, are shown to be resilient agains t a passive
adversary which tries to infer the training data only from the model parameters. However, recently, it has
been shown that the existing collaborative learning techniques are vulnerable to an active adversary that runs a GAN
attack during the learning phase. In this work, we propose a novel key-based col laborative learning technique that is
resilient against such GAN attacks. For this purpose, we present a collaborative learning formulation in which class scores
are protected by class-specific keys, and therefore, prevents a GAN attack. We a lso show that
very high dimensional class-specific keys can be utilized to improve robustness against attacks, without increasing the model complexity.
Our experimental results on two popular datasets, MNIST and AT&T Olivetti Faces, demonstrate the effectiveness of the proposed technique
against the GAN attack. To the best of our knowledge, the proposed approach is t he first collaborative learning
formulation that effectively tackles an active adversary, and, unlike model corr uption or differential privacy formulations,
our approach does not inherently feature a trade-off between model accuracy and data privacy.
****

Comparison of Paragram and GloVe Results for Similarity Benchmarks
Jakub Dutkiewicz,Czes■aw J■drzejek
Distributional Semantics Models(DSM) derive word space from linguistic items
in context. Meaning is obtained by defining a distance measure between vectors
corresponding to lexical entities. Such vectors present several problems. This
work concentrates on quality of word embeddings, improvement of word embedding
vectors, applicability of a novel similarity metric used ‘on top’ of the
word embeddings. In this paper we provide comparison between two methods
for post process improvements to the baseline DSM vectors. The counter-fitting
method which enforces antonymy and synonymy constraints into the Paragram
vector space representations recently showed improvement in the vectors’ capabil
ity
for judging semantic similarity. The second method is our novel RESM
method applied to GloVe baseline vectors. By applying the hubness reduction
method, implementing relational knowledge into the model by retrofitting synonym
s
and providing a new ranking similarity definition RESM that gives maximum
weight to the top vector component values we equal the results for the ESL
and TOEFL sets in comparison with our calculations using the Paragram and Paragr
am
+ Counter-fitting methods. For SIMLEX-999 gold standard since we cannot
use the RESM the results using GloVe and PPDB are significantly worse compared
to Paragram. Apparently, counter-fitting corrects hubness. The Paragram
or our cosine retrofitting method are state-of-the-art results for the SIMLEX-99
9
gold standard. They are 0.2 better for SIMLEX-999 than word2vec with sense
de-conflation (that was announced to be state-of the-art method for less reliabl
e
gold standards). Apparently relational knowledge and counter-fitting is more imp
ortant
for judging semantic similarity than sense determination for words. It is to
be mentioned, though that Paragram hyperparameters are fitted to SIMLEX-999
results. The lesson is that many corrections to word embeddings are necessary
and methods with more parameters and hyperparameters perform better.

****

Statestream: A toolbox to explore layerwise-parallel deep neural networks
Volker Fischer
Building deep neural networks to control autonomous agents which have to interac
t in real-time with the physical world, such as robots or automotive vehicles, r
equires a seamless integration of time into a network’s architecture. The centra
l question of this work is, how the temporal nature of reality should be reflect
ed in the execution of a deep neural network and its components. Most artificial
deep neural networks are partitioned into a directed graph of connected modules
or layers and the layers themselves consist of elemental building blocks, such
as single units. For most deep neural networks, all units of a layer are process
ed synchronously and in parallel, but layers themselves are processed in a seque
ntial manner. In contrast, all elements of a biological neural network are proce
ssed in parallel. In this paper, we define a class of networks between these two
extreme cases. These networks are executed in a streaming or synchronous layerw
ise-parallel manner, unlocking the layers of such networks for parallel processi
ng. Compared to the standard layerwise-sequential deep networks, these new layer
wise-parallel networks show a fundamentally different temporal behavior and flow
of information, especially for networks with skip or recurrent connections. We
argue that layerwise-parallel deep networks are better suited for future challen
ges of deep neural network design, such as large functional modularized and/or r
ecurrent architectures as well as networks allocating different network capaciti
es dependent on current stimulus and/or task complexity. We layout basic propert
ies and discuss major challenges for layerwise-parallel networks. Additionally,
we provide a toolbox to design, train, evaluate, and online-interact with layerw

ise-parallel networks.
****
Optimal transport maps for distribution preserving operations on latent spaces of Generative Models
Eirikur Agustsson,Alexander Sage,Radu Timofte,Luc Van Gool
Generative models such as Variational Auto Encoders (VAEs) and Generative Adversarial Networks (GANs) are typically trained for a fixed prior distribution in the latent space, such as uniform or Gaussian.
After a trained model is obtained, one can sample the Generator in various forms for exploration and understanding, such as interpolating between two samples, sampling in the vicinity of a sample or exploring differences between a pair of samples applied to a third sample.
In this paper, we show that the latent space operations used in the literature so far induce a distribution mismatch between the resulting outputs and the prior distribution the model was trained on. To address this, we propose to use distribution matching transport maps to ensure that such latent space operations preserve the prior distribution, while minimally modifying the original operation.
Our experimental results validate that the proposed operations give higher quality samples compared to the original operations.
****
Tandem Blocks in Deep Convolutional Neural Networks
Chris Hettinger,Tanner Christensen,Jeff Humpherys,Tyler J Jarvis
Due to the success of residual networks (resnets) and related architectures, shortcut connections have quickly become standard tools for building convolutional neural networks. The explanations in the literature for the apparent effectiveness of shortcuts are varied and often contradictory. We hypothesize that shortcuts work primarily because they act as linear counterparts to nonlinear layers. We test this hypothesis by using several variations on the standard residual block, with different types of linear connections, to build small (100k--1.2M parameter) image classification networks. Our experiments show that other kinds of linear connections can be even more effective than the identity shortcuts. Our results also suggest that the best type of linear connection for a given application may depend on both network width and depth.
****
Smooth Loss Functions for Deep Top-k Classification
Leonard Berrada,Andrew Zisserman,M. Pawan Kumar
The top-k error is a common measure of performance in machine learning and computer vision. In practice, top-k classification is typically performed with deep neural networks trained with the cross-entropy loss. Theoretical results indeed suggest that cross-entropy is an optimal learning objective for such a task in the limit of infinite data. In the context of limited and noisy data however, the use of a loss function that is specifically designed for top-k classification can bring significant improvements.
Our empirical evidence suggests that the loss function must be smooth and have non-sparse gradients in order to work well with deep neural networks. Consequently, we introduce a family of smoothed loss functions that are suited to top-k optimization via deep learning. The widely used cross-entropy is a special case of our family. Evaluating our smooth loss functions is computationally challenging: a na{\"i}ve algorithm would require $\mathcal{O}(\binom{n}{k})$ operations, where n is the number of classes. Thanks to a connection to polynomial algebra and a divide-and-conquer approach, we provide an algorithm with a time complexity of $\mathcal{O}(k n)$. Furthermore, we present a novel approximation to obtain fast and stable algorithms on GPUs with single floating point precision. We compare the performance of the cross-entropy loss and our margin-based losses in various regimes of noise and data size, for the predominant use case of $k=5$. Our investigation reveals that our loss is more robust to noise and overfitting than cross-entropy.
****
Entropy-SGD optimizes the prior of a PAC-Bayes bound: Data-dependent PAC-Bayes priors via differential privacy

Gintare Karolina Dziugaite,Daniel M. Roy
We show that Entropy-SGD (Chaudhari et al., 2017), when viewed as a learning algorithm, optimizes a PAC-Bayes bound on the risk of a Gibbs (posterior) classifier, i.e., a randomized classifier obtained by a risk-sensitive perturbation of the weights of a learned classifier. Entropy-SGD works by optimizing the bound's prior, violating the hypothesis of the PAC-Bayes theorem that the prior is chosen independently of the data. Indeed, available implementations of Entropy-SGD rapidly obtain zero training error on random labels and the same holds of the Gibbs posterior. In order to obtain a valid generalization bound, we show that an ε-differentially private prior yields a valid PAC-Bayes bound, a straightforward consequence of results connecting generalization with differential privacy. Using stochastic gradient Langevin dynamics (SGLD) to approximate the well-known exponential release mechanism, we observe that generalization error on MNIST (measured on held out data) falls within the (empirically nonvacuous) bounds computed under the assumption that SGLD produces perfect samples. In particular, Entropy-SGLD can be configured to yield relatively tight generalization bounds and still fit real labels, although these same settings do not obtain state-of-the-art performance.
****
Minimal-Entropy Correlation Alignment for Unsupervised Deep Domain Adaptation
Pietro Morerio,Jacopo Cavazza,Vittorio Murino
In this work, we face the problem of unsupervised domain adaptation with a novel deep learning approach which leverages our finding that entropy minimization is induced by the optimal alignment of second order statistics between source and target domains. We formally demonstrate this hypothesis and, aiming at achieving an optimal alignment in practical cases, we adopt a more principled strategy which, differently from the current Euclidean approaches, deploys alignment along geodesics. Our pipeline can be implemented by adding to the standard classification loss (on the labeled source domain), a source-to-target regularizer that is weighted in an unsupervised and data-driven fashion. We provide extensive experiments to assess the superiority of our framework on standard domain and modality adaptation benchmarks.
****
Automatic Parameter Tying in Neural Networks
Yibo Yang,Nicholas Ruozzi,Vibhav Gogate
Recently, there has been growing interest in methods that perform neural network compression, namely techniques that attempt to substantially reduce the size of a neural network without significant reduction in performance. However, most existing methods are post-processing approaches in that they take a learned neural network as input and output a compressed network by either forcing several parameters to take the same value (parameter tying via quantization) or pruning irrelevant edges (pruning) or both. In this paper, we propose a novel algorithm that jointly learns and compresses a neural network. The key idea in our approach is to change the optimization criteria by adding k independent Gaussian priors over the parameters and a sparsity penalty. We show that our approach is easy to implement using existing neural network libraries, generalizes L1 and L2 regularization and elegantly enforces parameter tying as well as pruning constraints. Experimentally, we demonstrate that our new algorithm yields state-of-the-art compression on several standard benchmarks with minimal loss in accuracy while requiring little to no hyperparameter tuning as compared with related, competing approaches.
****
PixelNN: Example-based Image Synthesis
Aayush Bansal,Yaser Sheikh,Deva Ramanan
We present a simple nearest-neighbor (NN) approach that synthesizes high-frequency photorealistic images from an ``incomplete'' signal such as a low-resolution image, a surface normal map, or edges. Current state-of-the-art deep generative models designed for such conditional image synthesis lack two important things: (1) they are unable to generate a large set of diverse outputs, due to the mode collapse problem. (2) they are not interpretable, making it difficult to control

the synthesized output. We demonstrate that NN approaches potentially address such limitations, but suffer in accuracy on small datasets. We design a simple pipeline that combines the best of both worlds: the first stage uses a convolutional neural network (CNN) to map the input to a (overly-smoothed) image, and the second stage uses a pixel-wise nearest neighbor method to map the smoothed output to multiple high-quality, high-frequency outputs in a controllable manner. Importantly, pixel-wise matching allows our method to compose novel high-frequency content by cutting-and-pasting pixels from different training exemplars. We demonstrate our approach for various input modalities, and for various domains ranging from human faces, pets, shoes, and handbags.
****

Adversarial reading networks for machine comprehension

Quentin Grail,Julien Perez

Machine reading has recently shown remarkable progress thanks to differentiable reasoning models. In this context, End-to-End trainable Memory Networks (MemN2N) have demonstrated promising performance on simple natural language based reasoning tasks such as factual reasoning and basic deduction. However, the task of machine comprehension is currently bounded to a supervised setting and available question answering dataset. In this paper we explore the paradigm of adversarial learning and self-play for the task of machine reading comprehension.

Inspired by the successful propositions in the domain of game learning, we present a novel approach of training for this task that is based on the definition of a coupled attention-based memory model. On one hand, a reader network is in charge of finding answers regarding a passage of text and a question. On the other hand, a narrator network is in charge of obfuscating spans of text in order to minimize the probability of success of the reader. We experimented the model on several question-answering corpora. The proposed learning paradigm and associated models present encouraging results.
****

DNN Model Compression Under Accuracy Constraints

Soroosh Khoram,Jing Li

The growing interest to implement Deep Neural Networks (DNNs) on resource-bound hardware has motivated innovation of compression algorithms. Using these algorithms, DNN model sizes can be substantially reduced, with little to no accuracy degradation. This is achieved by either eliminating components from the model, or penalizing complexity during training. While both approaches demonstrate considerable compressions, the former often ignores the loss function during compression while the later produces unpredictable compressions. In this paper, we propose a technique that directly minimizes both the model complexity and the changes in the loss function. In this technique, we formulate compression as a constrained optimization problem, and then present a solution for it. We will show that using this technique, we can achieve competitive results.
****

Autonomous Vehicle Fleet Coordination With Deep Reinforcement Learning

Cane Punma

Autonomous vehicles are becoming more common in city transportation. Companies will begin to find a need to teach these vehicles smart city fleet coordination. Currently, simulation based modeling along with hand coded rules dictate the decision making of these autonomous vehicles. We believe that complex intelligent behavior can be learned by these agents through Reinforcement Learning.In this paper, we discuss our work for solving this system by adapting the Deep Q-Learning (DQN) model to the multi-agent setting. Our approach applies deep reinforcement learning by combining convolutional neural networks with DQN to teach agents to fulfill customer demand in an environment that is partially observ-able to them. We also demonstrate how to utilize transfer learning to teach agents to balance multiple objectives such as navigating to a charging station when its en-er

gy level is low. The two evaluations presented show that our solution has shown hat we are successfully able to teach agents cooperation policies while balancing multiple objectives.
****
Deep Learning Inferences with Hybrid Homomorphic Encryption
Anthony Meehan,Ryan K L Ko,Geoff Holmes
When deep learning is applied to sensitive data sets, many privacy-related implementation issues arise. These issues are especially evident in the healthcare, finance, law and government industries. Homomorphic encryption could allow a server to make inferences on inputs encrypted by a client, but to our best knowledge, there has been no complete implementation of common deep learning operations, for arbitrary model depths, using homomorphic encryption. This paper demonstrates a novel approach, efficiently implementing many deep learning functions with bootstrapped homomorphic encryption. As part of our implementation, we demonstrate Single and Multi-Layer Neural Networks, for the Wisconsin Breast Cancer dataset, as well as a Convolutional Neural Network for MNIST. Our results give promising directions for privacy-preserving representation learning, and the return of data control to users.

****
Trust-PCL: An Off-Policy Trust Region Method for Continuous Control
Ofir Nachum,Mohammad Norouzi,Kelvin Xu,Dale Schuurmans
Trust region methods, such as TRPO, are often used to stabilize policy optimization algorithms in reinforcement learning (RL). While current trust region strategies are effective for continuous control, they typically require a large amount of on-policy interaction with the environment. To address this problem, we propose an off-policy trust region method, Trust-PCL, which exploits an observation that the optimal policy and state values of a maximum reward objective with a relative-entropy regularizer satisfy a set of multi-step pathwise consistencies along any path. The introduction of relative entropy regularization allows Trust-PCL to maintain optimization stability while exploiting off-policy data to improve sample efficiency. When evaluated on a number of continuous control tasks, Trust-PCL significantly improves the solution quality and sample efficiency of TRPO.
****
Graph2Seq: Scalable Learning Dynamics for Graphs
Shaileshh Bojja Venkatakrishnan,Mohammad Alizadeh,Pramod Viswanath
Neural networks are increasingly used as a general purpose approach to learning algorithms over graph structured data. However, techniques for representing graphs as real-valued vectors are still in their infancy. Recent works have proposed several approaches (e.g., graph convolutional networks), but as we show in this paper, these methods have difficulty generalizing to large graphs. In this paper we propose Graph2Seq, an embedding framework that represents graphs as an infinite time-series. By not limiting the representation to a fixed dimension, Graph2Seq naturally scales to graphs of arbitrary size. Moreover, through analysis of a formal computational model we show that an unbounded sequence is necessary for scalability. Graph2Seq is also reversible, allowing full recovery of the graph structure from the sequence. Experimental evaluations of Graph2Seq on a variety of combinatorial optimization problems show strong generalization and strict improvement over state of the art.
****
Residual Gated Graph ConvNets
Xavier Bresson,Thomas Laurent
Graph-structured data such as social networks, functional brain networks, gene regulatory networks, communications networks have brought the interest in generalizing deep learning techniques to graph domains. In this paper, we are intereste d to design neural networks for graphs with variable length in order to solve learning problems such as vertex classification, graph classification, graph regression, and graph generative tasks. Most existing works have focused on recurrent

neural networks (RNNs) to learn meaningful representations of graphs, and more recently new convolutional neural networks (ConvNets) have been introduced. In this work, we want to compare rigorously these two fundamental families of architectures to solve graph learning tasks. We review existing graph RNN and ConvNet architectures, and propose natural extension of LSTM and ConvNet to graphs with arbitrary size. Then, we design a set of analytically controlled experiments on two basic graph problems, i.e. subgraph matching and graph clustering, to test the different architectures. Numerical results show that the proposed graph ConvNets are 3-17% more accurate and 1.5-4x faster than graph RNNs. Graph ConvNets are also 36% more accurate than variational (non-learning) techniques. Finally, the most effective graph ConvNet architecture uses gated edges and residuality. Residuality plays an essential role to learn multi-layer architectures as they provide a 10% gain of performance.
****

Neural Clustering By Predicting And Copying Noise

Sam Coope,Andrej Zukov-Gregoric,Yoram Bachrach

We propose a neural clustering model that jointly learns both latent features and how they cluster. Unlike similar methods our model does not require a predefined number of clusters. Using a supervised approach, we agglomerate latent features towards randomly sampled targets within the same space whilst progressively removing the targets until we are left with only targets which represent cluster centroids. To show the behavior of our model across different modalities we apply our model on both text and image data and very competitive results on MNIST. Finally, we also provide results against baseline models for fashion-MNIST, the 20 newsgroups dataset, and a Twitter dataset we ourselves create.
****

Stochastic Activation Pruning for Robust Adversarial Defense

Guneet S. Dhillon,Kamyar Azizzadenesheli,Zachary C. Lipton,Jeremy D. Bernstein,Jean Kossaifi,Aran Khanna,Animashree Anandkumar

Neural networks are known to be vulnerable to adversarial examples. Carefully chosen perturbations to real images, while imperceptible to humans, induce misclassification and threaten the reliability of deep learning systems in the wild. To guard against adversarial examples, we take inspiration from game theory and cast the problem as a minimax zero-sum game between the adversary and the model. In general, for such games, the optimal strategy for both players requires a stochastic policy, also known as a mixed strategy. In this light, we propose Stochastic Activation Pruning (SAP), a mixed strategy for adversarial defense. SAP prunes a random subset of activations (preferentially pruning those with smaller magnitude) and scales up the survivors to compensate. We can apply SAP to pretrained networks, including adversarially trained models, without fine-tuning, providing robustness against adversarial examples. Experiments demonstrate that SAP confers robustness against attacks, increasing accuracy and preserving calibration.
****

Learning to Optimize Neural Nets

Ke Li,Jitendra Malik

Learning to Optimize is a recently proposed framework for learning optimization algorithms using reinforcement learning. In this paper, we explore learning an optimization algorithm for training shallow neural nets. Such high-dimensional stochastic optimization problems present interesting challenges for existing reinforcement learning algorithms. We develop an extension that is suited to learning optimization algorithms in this setting and demonstrate that the learned optimization algorithm consistently outperforms other known optimization algorithms even on unseen tasks and is robust to changes in stochasticity of gradients and the neural net architecture. More specifically, we show that an optimization algorithm trained with the proposed method on the problem of training a neural net on MNIST generalizes to the problems of training neural nets on the Toronto Faces Dataset, CIFAR-10 and CIFAR-100.
****

Generative Discovery of Relational Medical Entity Pairs

Chenwei Zhang,Yaliang Li,Nan Du,Wei Fan,Philip S. Yu

Online healthcare services can provide the general public with ubiquitous access to medical knowledge and reduce the information access cost for both individuals and societies. To promote these benefits, it is desired to effectively expand the scale of high-quality yet novel relational medical entity pairs that embody rich medical knowledge in a structured form. To fulfill this goal, we introduce a generative model called Conditional Relationship Variational Autoencoder (CRVAE), which can discover meaningful and novel relational medical entity pairs without the requirement of additional external knowledge. Rather than discriminatively identifying the relationship between two given medical entities in a free-text corpus, we directly model and understand medical relationships from diversely expressed medical entity pairs. The proposed model introduces the generative modeling capacity of variational autoencoder to entity pairs, and has the ability to discover new relational medical entity pairs solely based on the existing entity pairs. Beside entity pairs, relationship-enhanced entity representations are obtained as another appealing benefit of the proposed method. Both quantitative and qualitative evaluations on real-world medical datasets demonstrate the effectiveness of the proposed method in generating relational medical entity pairs that are meaningful and novel.
****
Multi-Advisor Reinforcement Learning
Romain Laroche,Mehdi Fatemi,Joshua Romoff,Harm van Seijen
We consider tackling a single-agent RL problem by distributing it to n learners. These learners, called advisors, endeavour to solve the problem from a different focus. Their advice, taking the form of action values, is then communicated to an aggregator, which is in control of the system. We show that the local planning method for the advisors is critical and that none of the ones found in the literature is flawless: the \textit{egocentric} planning overestimates values of states where the other advisors disagree, and the \textit{agnostic} planning is inefficient around danger zones. We introduce a novel approach called \textit{empathic} and discuss its theoretical aspects. We empirically examine and validate our theoretical findings on a fruit collection task.
****
Learning Sparse Latent Representations with the Deep Copula Information Bottleneck
Aleksander Wieczorek*,Mario Wieser*,Damian Murezzan,Volker Roth
Deep latent variable models are powerful tools for representation learning. In this paper, we adopt the deep information bottleneck model, identify its shortcomings and propose a model that circumvents them. To this end, we apply a copula transformation which, by restoring the invariance properties of the information bottleneck method, leads to disentanglement of the features in the latent space. Building on that, we show how this transformation translates to sparsity of the latent space in the new model. We evaluate our method on artificial and real data.
****
Joint autoencoders: a flexible meta-learning framework
Baruch Epstein,Ron Meir,Tomer Michaeli
The incorporation of prior knowledge into learning is essential in achieving good performance based on small noisy samples. Such knowledge is often incorporated through the availability of related data arising from domains and tasks similar to the one of current interest. Ideally one would like to allow both the data for the current task and for previous related tasks to self-organize the learning system in such a way that commonalities and differences between the tasks are learned in a data-driven fashion. We develop a framework for learning multiple tasks simultaneously, based on sharing features that are common to all tasks, achieved through the use of a modular deep feedforward neural network consisting of shared branches, dealing with the common features of all tasks, and private branches, learning the specific unique aspects of each task. Once an appropriate weight sharing architecture has been established, learning takes place through standard algorithms for feedforward networks, e.g., stochastic gradient descent and its variations. The method deals with meta-learning (such as domain adaptation,

transfer and multi-task learning) in a unified fashion, and can easily deal with data arising from different types of sources. Numerical experiments demonstrate the effectiveness of learning in domain adaptation and transfer learning setups, and provide evidence for the flexible and task-oriented representations arising in the network.
****
Analyzing and Exploiting NARX Recurrent Neural Networks for Long-Term Dependencies
Robert DiPietro,Christian Rupprecht,Nassir Navab,Gregory D. Hager
Recurrent neural networks (RNNs) have achieved state-of-the-art performance on many diverse tasks, from machine translation to surgical activity recognition, yet training RNNs to capture long-term dependencies remains difficult. To date, the vast majority of successful RNN architectures alleviate this problem using nearly-additive connections between states, as introduced by long short-term memory (LSTM). We take an orthogonal approach and introduce MIST RNNs, a NARX RNN architecture that allows direct connections from the very distant past. We show that MIST RNNs 1) exhibit superior vanishing-gradient properties in comparison to LSTM and previously-proposed NARX RNNs; 2) are far more efficient than previously-proposed NARX RNN architectures, requiring even fewer computations than LSTM; and 3) improve performance substantially over LSTM and Clockwork RNNs on tasks requiring very long-term dependencies.
****
Flexible Prior Distributions for Deep Generative Models
Yannic Kilcher,Aurelien Lucchi,Thomas Hofmann
We consider the problem of training generative models with deep neural networks as generators, i.e. to map latent codes to data points. Whereas the dominant paradigm combines simple priors over codes with complex deterministic models, we argue that it might be advantageous to use more flexible code distributions. We demonstrate how these distributions can be induced directly from the data. The benefits include: more powerful generative models, better modeling of latent structure and explicit control of the degree of generalization.
****
Meta-Learning for Semi-Supervised Few-Shot Classification
Mengye Ren,Eleni Triantafillou,Sachin Ravi,Jake Snell,Kevin Swersky,Joshua B. Tenenbaum,Hugo Larochelle,Richard S. Zemel
In few-shot classification, we are interested in learning algorithms that train a classifier from only a handful of labeled examples. Recent progress in few-shot classification has featured meta-learning, in which a parameterized model for a learning algorithm is defined and trained on episodes representing different classification problems, each with a small labeled training set and its corresponding test set. In this work, we advance this few-shot classification paradigm towards a scenario where unlabeled examples are also available within each episode. We consider two situations: one where all unlabeled examples are assumed to belong to the same set of classes as the labeled examples of the episode, as well as the more challenging situation where examples from other distractor classes are also provided. To address this paradigm, we propose novel extensions of Prototypical Networks (Snell et al., 2017) that are augmented with the ability to use unlabeled examples when producing prototypes. These models are trained in an end-to-end way on episodes, to learn to leverage the unlabeled examples successfully. We evaluate these methods on versions of the Omniglot and miniImageNet benchmarks, adapted to this new framework augmented with unlabeled examples. We also propose a new split of ImageNet, consisting of a large set of classes, with a hierarchical structure. Our experiments confirm that our Prototypical Networks can learn to improve their predictions due to unlabeled examples, much like a semi-supervised algorithm would.
****
Prediction Under Uncertainty with Error Encoding Networks
Mikael Henaff,Junbo Zhao,Yann Lecun
In this work we introduce a new framework for performing temporal predictions in the presence of uncertainty. It is based on a simple idea of disentangling co

m-
ponents of the future state which are predictable from those which are inherentl
y
unpredictable, and encoding the unpredictable components into a low-dimensional
latent variable which is fed into the forward model. Our method uses a simple su
-
pervised training objective which is fast and easy to train. We evaluate it in t
he
context of video prediction on multiple datasets and show that it is able to con
si-
tently generate diverse predictions without the need for alternating minimizatio
n
over a latent space or adversarial training.
****

Stochastic Variational Video Prediction
Mohammad Babaeizadeh,Chelsea Finn,Dumitru Erhan,Roy H. Campbell,Sergey Levine
Predicting the future in real-world settings, particularly from raw sensory observations such as images, is exceptionally challenging. Real-world events can be stochastic and unpredictable, and the high dimensionality and complexity of natural images requires the predictive model to build an intricate understanding of the natural world. Many existing methods tackle this problem by making simplifying assumptions about the environment. One common assumption is that the outcome is deterministic and there is only one plausible future. This can lead to low-quality predictions in real-world settings with stochastic dynamics. In this paper, we develop a stochastic variational video prediction (SV2P) method that predicts a different possible future for each sample of its latent variables. To the best of our knowledge, our model is the first to provide effective stochastic multi-frame prediction for real-world video. We demonstrate the capability of the proposed method in predicting detailed future frames of videos on multiple real-world datasets, both action-free and action-conditioned. We find that our proposed method produces substantially improved video predictions when compared to the same model without stochasticity, and to other stochastic video prediction methods. Our SV2P implementation will be open sourced upon publication.
****

Data Augmentation by Pairing Samples for Images Classification
Hiroshi Inoue
Data augmentation is a widely used technique in many machine learning tasks, such as image classification, to virtually enlarge the training dataset size and avoid overfitting. Traditional data augmentation techniques for image classification tasks create new samples from the original training data by, for example, flipping, distorting, adding a small amount of noise to, or cropping a patch from an original image. In this paper, we introduce a simple but surprisingly effective data augmentation technique for image classification tasks. With our technique, named SamplePairing, we synthesize a new sample from one image by overlaying another image randomly chosen from the training data (i.e., taking an average of two images for each pixel). By using two images randomly selected from the training set, we can generate N^2 new samples from N training samples. This simple data augmentation technique significantly improved classification accuracy for all the tested datasets; for example, the top-1 error rate was reduced from 33.5% to 29.0% for the ILSVRC 2012 dataset with GoogLeNet and from 8.22% to 6.93% in the CIFAR-10 dataset. We also show that our SamplePairing technique largely improved accuracy when the number of samples in the training set was very small. Therefore, our technique is more valuable for tasks with a limited amount of training data, such as medical imaging tasks.

****

Learning Deep ResNet Blocks Sequentially using Boosting Theory
Furong Huang,Jordan T. Ash,John Langford,Robert E. Schapire
We prove a multiclass boosting theory for the ResNet architectures which simultaneously creates a new technique for multiclass boosting and provides a new algor

ithm for ResNet-style architectures. Our proposed training algorithm, BoostResNet, is particularly suitable in non-differentiable architectures. Our method only requires the relatively inexpensive sequential training of T "shallow ResNets". We prove that the training error decays exponentially with the depth T if the weak module classifiers that we train perform slightly better than some weak baseline. In other words, we propose a weak learning condition and prove a boosting theory for ResNet under the weak learning condition. A generalization error bound based on margin theory is proved and suggests that ResNet could be resistant to overfitting using a network with l_1 norm bounded weights.
****
Multimodal Sentiment Analysis To Explore the Structure of Emotions
Anthony Hu,Seth Flaxman
We propose a novel approach to multimodal sentiment analysis using deep neural networks combining visual recognition and natural language processing. Our goal is different than the standard sentiment analysis goal of predicting whether a sentence expresses positive or negative sentiment; instead, we aim to infer the latent emotional state of the user. Thus, we focus on predicting the emotion word tags attached by users to their Tumblr posts, treating these as “self-reported emotions.”
We demonstrate that our multimodal model combining both text and image features outperforms separate models based solely on either images or text. Our model’s results are interpretable, automatically yielding sensible word lists associated with emotions. We explore the structure of emotions implied by our model and compare it to what has been posited in the psychology literature, and validate our model on a set of images that have been used in psychology studies. Finally, our work also provides a useful tool for the growing academic study of images—both photographs and memes—on social networks.
****
An efficient framework for learning sentence representations
Lajanugen Logeswaran,Honglak Lee
In this work we propose a simple and efficient framework for learning sentence representations from unlabelled data. Drawing inspiration from the distributional hypothesis and recent work on learning sentence representations, we reformulate the problem of predicting the context in which a sentence appears as a classification problem. Given a sentence and the context in which it appears, a classifier distinguishes context sentences from other contrastive sentences based on their vector representations. This allows us to efficiently learn different types of encoding functions, and we show that the model learns high-quality sentence representations. We demonstrate that our sentence representations outperform state-of-the-art unsupervised and supervised representation learning methods on several downstream NLP tasks that involve understanding sentence semantics while achieving an order of magnitude speedup in training time.
****
Learning what to learn in a neural program
Richard Shin,Dawn Song
Learning programs with neural networks is a challenging task, addressed by a long line of existing work. It is difficult to learn neural networks which will generalize to problem instances that are much larger than those used during training. Furthermore, even when the learned neural program empirically works on all test inputs, we cannot verify that it will work on every possible input. Recent work has shown that it is possible to address these issues by using recursion in the Neural Programmer-Interpreter, but this technique requires a verification set which is difficult to construct without knowledge of the internals of the oracle used to generate training data. In this work, we show how to automatically build such a verification set, which can also be directly used for training. By int

eractively querying an oracle, we can construct this set with minimal additional knowledge about the oracle. We empirically demonstrate that our method allows automated learning and verification of a recursive NPI program with provably perfect generalization.

****

Learning Independent Features with Adversarial Nets for Non-linear ICA

Philemon Brakel,Yoshua Bengio

Reliable measures of statistical dependence could potentially be useful tools for learning independent features and performing tasks like source separation using Independent Component Analysis (ICA). Unfortunately, many of such measures, like the mutual information, are hard to estimate and optimize directly. We propose to learn independent features with adversarial objectives (Goodfellow et al. 2014, Arjovsky et al. 2017) which optimize such measures implicitly. These objectives compare samples from the joint distribution and the product of the marginals without the need to compute any probability densities. We also propose two methods for obtaining samples from the product of the marginals using either a simple resampling trick or a separate parametric distribution. Our experiments show that this strategy can easily be applied to different types of model architectures and solve both linear and non-linear ICA problems.

****

Exploring Deep Recurrent Models with Reinforcement Learning for Molecule Design

Daniel Neil,Marwin Segler,Laura Guasch,Mohamed Ahmed,Dean Plumbley,Matthew Sellwood,Nathan Brown

The design of small molecules with bespoke properties is of central importance to drug discovery. However significant challenges yet remain for computational methods, despite recent advances such as deep recurrent networks and reinforcement learning strategies for sequence generation, and it can be difficult to compare results across different works. This work proposes 19 benchmarks selected by subject experts, expands smaller datasets previously used to approximately 1.1 million training molecules, and explores how to apply new reinforcement learning techniques effectively for molecular design. The benchmarks here, built as OpenAI Gym environments, will be open-sourced to encourage innovation in molecular design algorithms and to enable usage by those without a background in chemistry. Finally, this work explores recent development in reinforcement-learning methods with excellent sample complexity (the A2C and PPO algorithms) and investigates their behavior in molecular generation, demonstrating significant performance gains compared to standard reinforcement learning techniques.

****

Generating Adversarial Examples with Adversarial Networks

Chaowei Xiao,Bo Li,Jun-Yan Zhu,Warren He,Mingyan Liu,Dawn Song

Deep neural networks (DNNs) have been found to be vulnerable to adversarial examples resulting from adding small-magnitude perturbations to inputs. Such adversarial examples can mislead DNNs to produce adversary-selected results.
Different attack strategies have been proposed to generate adversarial examples, but how to produce them with high perceptual quality and more efficiently requires more research efforts.
In this paper, we propose AdvGAN to generate adversarial examples with generative adversarial networks (GANs), which can learn and approximate the distribution of original instances.
For AdvGAN, once the generator is trained, it can generate adversarial perturbations efficiently for any instance, so as to potentially accelerate adversarial training as defenses.
We apply AdvGAN in both semi-whitebox and black-box attack settings. In semi-whitebox attacks, there is no need to access the original target model after the generator is trained, in contrast to traditional white-box attacks. In black-box attacks, we dynamically train a distilled model for the black-box model and optimize the generator accordingly.
Adversarial examples generated by AdvGAN on different target models have high at

tack success rate under state-of-the-art defenses compared to other attacks. Our
 attack has placed the first with 92.76% accuracy on a public MNIST black-box a
ttack challenge.
****
Discriminative k-shot learning using probabilistic models
Matthias Bauer,Mateo Rojas-Carulla,Jakub Bart■omiej ■wi■tkowski,Bernhard Schölko
pf,Richard E. Turner
This paper introduces a probabilistic framework for k-shot image classification.
 The goal is to generalise from an initial large-scale classification task to a
separate task comprising new classes and small numbers of examples. The new app
roach not only leverages the feature-based representation learned by a neural ne
twork from the initial task (representational transfer), but also information ab
out the classes (concept transfer). The concept information is encapsulated in a
 probabilistic model for the final layer weights of the neural network which act
s as a prior for probabilistic k-shot learning. We show that even a simple proba
bilistic model achieves state-of-the-art on a standard k-shot learning dataset b
y a large margin. Moreover, it is able to accurately model uncertainty, leading
to well calibrated classifiers, and is easily extensible and flexible, unlike ma
ny recent approaches to k-shot learning.
****
On the importance of single directions for generalization
Ari S. Morcos,David G.T. Barrett,Neil C. Rabinowitz,Matthew Botvinick
Despite their ability to memorize large datasets, deep neural networks often ach
ieve good generalization performance. However, the differences between the learn
ed solutions of networks which generalize and those which do not remain unclear.
 Additionally, the tuning properties of single directions (defined as the activa
tion of a single unit or some linear combination of units in response to some in
put) have been highlighted, but their importance has not been evaluated. Here, w
e connect these lines of inquiry to demonstrate that a network’s reliance on sin
gle directions is a good predictor of its generalization performance, across net
works trained on datasets with different fractions of corrupted labels, across e
nsembles of networks trained on datasets with unmodified labels, across differen
t hyper- parameters, and over the course of training. While dropout only regular
izes this quantity up to a point, batch normalization implicitly discourages sin
gle direction reliance, in part by decreasing the class selectivity of individua
l units. Finally, we find that class selectivity is a poor predictor of task imp
ortance, suggesting not only that networks which generalize well minimize their
dependence on individual units by reducing their selectivity, but also that indi
vidually selective units may not be necessary for strong network performance.
****
Combining Symbolic Expressions and Black-box Function Evaluations in Neural Prog
rams
Forough Arabshahi,Sameer Singh,Animashree Anandkumar
Neural programming involves training neural networks to learn programs, mathemat
ics, or logic from data. Previous works have failed to achieve good generalizati
on performance, especially on problems and programs with high complexity or on l
arge domains. This is because they mostly rely either on black-box function eval
uations that do not capture the structure of the program, or on detailed executi
on traces that are expensive to obtain, and hence the training data has poor cov
erage of the domain under consideration. We present a novel framework that utili
zes black-box function evaluations, in conjunction with symbolic expressions tha
t define relationships between the given functions. We employ tree LSTMs to inco
rporate the structure of the symbolic expression trees. We use tree encoding for
 numbers present in function evaluation data, based on their decimal representat
ion. We present an evaluation benchmark for this task to demonstrate our propose
d model combines symbolic reasoning and function evaluation in a fruitful manner
, obtaining high accuracies in our experiments. Our framework generalizes signif
icantly better to expressions of higher depth and is able to fill partial equati
ons with valid completions.
****

Image Quality Assessment Techniques Improve Training and Evaluation of Energy-Based Generative Adversarial Networks
Michael O. Vertolli,Jim Davies
We propose a new, multi-component energy function for energy-based Generative Adversarial Networks (GANs) based on methods from the image quality assessment literature. Our approach expands on the Boundary Equilibrium Generative Adversarial Network (BEGAN) by outlining some of the short-comings of the original energy and loss functions. We address these short-comings by incorporating an l1 score, the Gradient Magnitude Similarity score, and a chrominance score into the new energy function. We then provide a set of systematic experiments that explore its hyper-parameters. We show that each of the energy function's components is able to represent a slightly different set of features, which require their own evaluation criteria to assess whether they have been adequately learned. We show that models using the new energy function are able to produce better image representations than the BEGAN model in predicted ways.
****

Deep Complex Networks
Chiheb Trabelsi,Olexa Bilaniuk,Ying Zhang,Dmitriy Serdyuk,Sandeep Subramanian,Joao Felipe Santos,Soroush Mehri,Negar Rostamzadeh,Yoshua Bengio,Christopher J Pal
At present, the vast majority of building blocks, techniques, and architectures for deep learning are based on real-valued operations and representations. However, recent work on recurrent neural networks and older fundamental theoretical analysis suggests that complex numbers could have a richer representational capacity and could also facilitate noise-robust memory retrieval mechanisms. Despite their attractive properties and potential for opening up entirely new neural architectures, complex-valued deep neural networks have been marginalized due to the absence of the building blocks required to design such models. In this work, we provide the key atomic components for complex-valued deep neural networks and apply them to convolutional feed-forward networks. More precisely, we rely on complex convolutions and present algorithms for complex batch-normalization, complex weight initialization strategies for complex-valued neural nets and we use them in experiments with end-to-end training schemes. We demonstrate that such complex-valued models are competitive with their real-valued counterparts. We test deep complex models on several computer vision tasks, on music transcription using the MusicNet dataset and on Speech spectrum prediction using TIMIT. We achieve state-of-the-art performance on these audio-related tasks.
****

Contextual memory bandit for pro-active dialog engagement
julien perez,Tomi Silander
An objective of pro-activity in dialog systems is to enhance the usability of conversational
agents by enabling them to initiate conversation on their own. While
dialog systems have become increasingly popular during the last couple of years,
current task oriented dialog systems are still mainly reactive and users tend to
initiate conversations. In this paper, we propose to introduce the paradigm of contextual
bandits as framework for pro-active dialog systems. Contextual bandits
have been the model of choice for the problem of reward maximization with partial
feedback since they fit well to the task description. As a second contribution,
we introduce and explore the notion of memory into this paradigm. We propose
two differentiable memory models that act as parts of the parametric reward estimation
function. The first one, Convolutional Selective Memory Networks, uses
a selection of past interactions as part of the decision support. The second model,
called Contextual Attentive Memory Network, implements a differentiable attention
mechanism over the past interactions of the agent. The goal is to generalize
the classic model of contextual bandits to settings where temporal information

needs to be incorporated and leveraged in a learnable manner. Finally, we illustrate
the usability and performance of our model for building a pro-active mobile assistant through an extensive set of experiments.
****

Policy Gradient For Multidimensional Action Spaces: Action Sampling and Entropy Bonus

Vuong Ho Quan,Yiming Zhang,Kenny Song,Xiao-Yue Gong,Keith W. Ross

In recent years deep reinforcement learning has been shown to be adept at solving sequential decision processes with high-dimensional state spaces such as in the Atari games. Many reinforcement learning problems, however, involve high-dimensional discrete action spaces as well as high-dimensional state spaces. In this paper, we develop a novel policy gradient methodology for the case of large multidimensional discrete action spaces. We propose two approaches for creating parameterized policies: LSTM parameterization and a Modified MDP (MMDP) giving rise to Feed-Forward Network (FFN) parameterization. Both of these approaches provide expressive models to which backpropagation can be applied for training. We then consider entropy bonus, which is typically added to the reward function to enhance exploration. In the case of high-dimensional action spaces, calculating the entropy and the gradient of the entropy requires enumerating all the actions in the action space and running forward and backpropagation for each action, which may be computationally infeasible. We develop several novel unbiased estimators for the entropy bonus and its gradient. Finally, we test our algorithms on two environments: a multi-hunter multi-rabbit grid game and a multi-agent multi-arm bandit problem.
****

A Flexible Approach to Automated RNN Architecture Generation

Martin Schrimpf,Stephen Merity,James Bradbury,Richard Socher

The process of designing neural architectures requires expert knowledge and extensive trial and error.
While automated architecture search may simplify these requirements, the recurrent neural network (RNN) architectures generated by existing methods are limited in both flexibility and components.
We propose a domain-specific language (DSL) for use in automated architecture search which can produce novel RNNs of arbitrary depth and width.
The DSL is flexible enough to define standard architectures such as the Gated Recurrent Unit and Long Short Term Memory and allows the introduction of non-standard RNN components such as trigonometric curves and layer normalization. Using two different candidate generation techniques, random search with a ranking function and reinforcement learning,
we explore the novel architectures produced by the RNN DSL for language modeling and machine translation domains.
The resulting architectures do not follow human intuition yet perform well on their targeted tasks, suggesting the space of usable RNN architectures is far larger than previously assumed.
****

TRL: Discriminative Hints for Scalable Reverse Curriculum Learning

Chen Wang,Xiangyu Chen,Zelin Ye,Jialu Wang,Ziruo Cai,Shixiang Gu,Cewu Lu

Deep reinforcement learning algorithms have proven successful in a variety of domains. However, tasks with sparse rewards remain challenging when the state space is large. Goal-oriented tasks are among the most typical problems in this domain, where a reward can only be received when the final goal is accomplished. In this work, we propose a potential solution to such problems with the introduction of an experience-based tendency reward mechanism, which provides the agent with additional hints based on a discriminative learning on past experiences during an automated reverse curriculum. This mechanism not only provides dense additional learning signals on what states lead to success, but also allows the agent to retain only this tendency reward instead of the whole histories of experience during multi-phase curriculum learning. We extensively study the advantages of our method on the standard sparse reward domains like Maze and Super Mario Bros a

nd show that our method performs more efficiently and robustly than prior approa ches in tasks with long time horizons and large state space. In addition, we dem onstrate that using an optional keyframe scheme with very small quantity of key states, our approach can solve difficult robot manipulation challenges directly from perception and sparse rewards.
****
Compressing Word Embeddings via Deep Compositional Code Learning
Raphael Shu,Hideki Nakayama
Natural language processing (NLP) models often require a massive number of param eters for word embeddings, resulting in a large storage or memory footprint. Dep loying neural NLP models to mobile devices requires compressing the word embeddi ngs without any significant sacrifices in performance. For this purpose, we prop ose to construct the embeddings with few basis vectors. For each word, the compo sition of basis vectors is determined by a hash code. To maximize the compressio n rate, we adopt the multi-codebook quantization approach instead of binary codi ng scheme. Each code is composed of multiple discrete numbers, such as (3, 2, 1, 8), where the value of each component is limited to a fixed range. We propose t o directly learn the discrete codes in an end-to-end neural network by applying the Gumbel-softmax trick. Experiments show the compression rate achieves 98% in a sentiment analysis task and 94% ~ 99% in machine translation tasks without per formance loss. In both tasks, the proposed method can improve the model performa nce by slightly lowering the compression rate. Compared to other approaches such as character-level segmentation, the proposed method is language-independent an d does not require modifications to the network architecture.
****
Viterbi-based Pruning for Sparse Matrix with Fixed and High Index Compression Ra tio
Dongsoo Lee,Daehyun Ahn,Taesu Kim,Pierce I. Chuang,Jae-Joon Kim
Weight pruning has proven to be an effective method in reducing the model size a nd computation cost while not sacrificing the model accuracy. Conventional spars e matrix formats, however, involve irregular index structures with large storage requirement and sequential reconstruction process, resulting in inefficient use of highly parallel computing resources. Hence, pruning is usually restricted to inference with a batch size of one, for which an efficient parallel matrix-vect or multiplication method exists. In this paper, a new class of sparse matrix rep resentation utilizing Viterbi algorithm that has a high, and more importantly, f ixed index compression ratio regardless of the pruning rate, is proposed. In thi s approach, numerous sparse matrix candidates are first generated by the Viterbi encoder, and then the one that aims to minimize the model accuracy degradation is selected by the Viterbi algorithm. The model pruning process based on the pro posed Viterbi encoder and Viterbi algorithm is highly parallelizable, and can be implemented efficiently in hardware to achieve low-energy, high-performance ind ex decoding process. Compared with the existing magnitude-based pruning methods, index data storage requirement can be further compressed by 85.2% in MNIST and 83.9% in AlexNet while achieving similar pruning rate. Even compared with the re lative index compression technique, our method can still reduce the index storag e requirement by 52.7% in MNIST and 35.5% in AlexNet.
****
Interpretable Counting for Visual Question Answering
Alexander Trott,Caiming Xiong,Richard Socher
Questions that require counting a variety of objects in images remain a major ch allenge in visual question answering (VQA). The most common approaches to VQA in volve either classifying answers based on fixed length representations of both t he image and question or summing fractional counts estimated from each section o f the image. In contrast, we treat counting as a sequential decision process and force our model to make discrete choices of what to count. Specifically, the mo del sequentially selects from detected objects and learns interactions between o bjects that influence subsequent selections. A distinction of our approach is it s intuitive and interpretable output, as discrete counts are automatically groun ded in the image. Furthermore, our method outperforms the state of the art archi

tecture for VQA on multiple metrics that evaluate counting.
****

NOVEL AND EFFECTIVE PARALLEL MIX-GENERATOR GENERATIVE ADVERSARIAL NETWORKS
Xia Xiao,Sanguthevar Rajasekaran
In this paper, we propose a mix-generator generative adversarial networks (PGAN) model that works in parallel by mixing multiple disjoint generators to approximate a complex real distribution. In our model, we propose an adjustment component that collects all the generated data points from the generators, learns the boundary between each pair of generators, and provides error to separate the support of each of the generated distributions. To overcome the instability in a multiplayer game, a shrinkage adjustment component method is introduced to gradually reduce the boundary between generators during the training procedure. To address the linearly growing training time problem in a multiple generators model, we propose a method to train the generators in parallel. This means that our work can be scaled up to large parallel computation frameworks. We present an efficient loss function for the discriminator, an effective adjustment component, and a suitable generator. We also show how to introduce the decay factor to stabilize the training procedure. We have performed extensive experiments on synthetic datasets, MNIST, and CIFAR-10. These experiments reveal that the error provided by the adjustment component could successfully separate the generated distributions and each of the generators can stably learn a part of the real distribution even if only a few modes are contained in the real distribution.
****

The Mutual Autoencoder: Controlling Information in Latent Code Representations
Mary Phuong,Max Welling,Nate Kushman,Ryota Tomioka,Sebastian Nowozin
Variational autoencoders (VAE) learn probabilistic latent variable models by optimizing a bound on the marginal likelihood of the observed data. Beyond providing a good density model a VAE model assigns to each data instance a latent code. In many applications, this latent code provides a useful high-level summary of the observation. However, the VAE may fail to learn a useful representation when the decoder family is very expressive. This is because maximum likelihood does not explicitly encourage useful representations and the latent variable is used only if it helps model the marginal distribution. This makes representation learning with VAEs unreliable. To address this issue, we propose a method for explicitly controlling the amount of information stored in the latent code. Our method can learn codes ranging from independent to nearly deterministic while benefiting from decoder capacity. Thus, we decouple the choice of decoder capacity and the latent code dimensionality from the amount of information stored in the code.

****

Semantically Decomposing the Latent Spaces of Generative Adversarial Networks
Chris Donahue,Zachary C. Lipton,Akshay Balsubramani,Julian McAuley
We propose a new algorithm for training generative adversarial networks to jointly learn latent codes for both identities (e.g. individual humans) and observations (e.g. specific photographs). In practice, this means that by fixing the identity portion of latent codes, we can generate diverse images of the same subject, and by fixing the observation portion we can traverse the manifold of subjects while maintaining contingent aspects such as lighting and pose. Our algorithm features a pairwise training scheme in which each sample from the generator consists of two images with a common identity code. Corresponding samples from the real dataset consist of two distinct photographs of the same subject. In order to fool the discriminator, the generator must produce images that are both photorealistic, distinct, and appear to depict the same person. We augment both the DCGAN and BEGAN approaches with Siamese discriminators to accommodate pairwise training. Experiments with human judges and an off-the-shelf face verification system demonstrate our algorithm’s ability to generate convincing, identity-matched photographs.
****

Adaptive Quantization of Neural Networks
Soroosh Khoram,Jing Li

Despite the state-of-the-art accuracy of Deep Neural Networks (DNN) in various classification problems, their deployment onto resource constrained edge computing devices remains challenging due to their large size and complexity. Several recent studies have reported remarkable results in reducing this complexity through quantization of DNN models. However, these studies usually do not consider the changes in the loss function when performing quantization, nor do they take the different importances of DNN model parameters to the accuracy into account. We address these issues in this paper by proposing a new method, called adaptive quantization, which simplifies a trained DNN model by finding a unique, optimal precision for each network parameter such that the increase in loss is minimized. The optimization problem at the core of this method iteratively uses the loss function gradient to determine an error margin for each parameter and assigns it a precision accordingly. Since this problem uses linear functions, it is computationally cheap and, as we will show, has a closed-form approximate solution. Experiments on MNIST, CIFAR, and SVHN datasets showed that the proposed method can achieve near or better than state-of-the-art reduction in model size with similar error rates. Furthermore, it can achieve compressions close to floating-point model compression methods without loss of accuracy.
****
UPS: optimizing Undirected Positive Sparse graph for neural graph filtering
Mikhail Yurochkin,Dung Thai,Hung Hai Bui,XuanLong Nguyen
In this work we propose a novel approach for learning graph representation of the data using gradients obtained via backpropagation. Next we build a neural network architecture compatible with our optimization approach and motivated by graph filtering in the vertex domain. We demonstrate that the learned graph has richer structure than often used nearest neighbors graphs constructed based on features similarity. Our experiments demonstrate that we can improve prediction quality for several convolution on graphs architectures, while others appeared to be insensitive to the input graph.
****
Rethinking the Smaller-Norm-Less-Informative Assumption in Channel Pruning of Convolution Layers
Jianbo Ye,Xin Lu,Zhe Lin,James Z. Wang
Model pruning has become a useful technique that improves the computational efficiency of deep learning, making it possible to deploy solutions in resource-limited scenarios. A widely-used practice in relevant work assumes that a smaller-norm parameter or feature plays a less informative role at the inference time. In this paper, we propose a channel pruning technique for accelerating the computations of deep convolutional neural networks (CNNs) that does not critically rely on this assumption. Instead, it focuses on direct simplification of the channel-to-channel computation graph of a CNN without the need of performing a computationally difficult and not-always-useful task of making high-dimensional tensors of CNN structured sparse. Our approach takes two stages: first to adopt an end-to-end stochastic training method that eventually forces the outputs of some channels to be constant, and then to prune those constant channels from the original neural network by adjusting the biases of their impacting layers such that the resulting compact model can be quickly fine-tuned. Our approach is mathematically appealing from an optimization perspective and easy to reproduce. We experimented our approach through several image learning benchmarks and demonstrate its interest- ing aspects and competitive performance.
****
GraphVAE: Towards Generation of Small Graphs Using Variational Autoencoders
Martin Simonovsky,Nikos Komodakis
Deep learning on graphs has become a popular research topic with many applications. However, past work has concentrated on learning graph embedding tasks only, which is in contrast with advances in generative models for images and text. Is it possible to transfer this progress to the domain of graphs? We propose to sidestep hurdles associated with linearization of such discrete structures by having a decoder output a probabilistic fully-connected graph of a predefined maximum size directly at once. Our method is formulated as a variational autoencoder. W

e evaluate on the challenging task of conditional molecule generation.
****

Learning to Select: Problem, Solution, and Applications
Heechang Ryu,Donghyun Kim,Hayong Shin
We propose a "Learning to Select" problem that selects the best among the flexib le size candidates. This makes decisions based not only on the properties of the candidate, but also on the environment in which they belong to. For example, jo b dispatching in the manufacturing factory is a typical "Learning to Select" pro blem. We propose Variable-Length CNN which combines the classification power usi ng hidden features from CNN and the idea of flexible input from Learning to Rank algorithms. This not only can handles flexible candidates using Dynamic Computa tion Graph, but also is computationally efficient because it only builds a netwo rk with the necessary sizes to fit the situation. We applied the algorithm to th e job dispatching problem which uses the dispatching log data obtained from the virtual fine-tuned factory. Our proposed algorithm shows considerably better per formance than other comparable algorithms.
****

Intrinsic Motivation and Automatic Curricula via Asymmetric Self-Play
Sainbayar Sukhbaatar,Zeming Lin,Ilya Kostrikov,Gabriel Synnaeve,Arthur Szlam,Rob Fergus
We describe a simple scheme that allows an agent to learn about its environment in an unsupervised manner. Our scheme pits two versions of the same agent, Alice and Bob, against one another. Alice proposes a task for Bob to complete; and th en Bob attempts to complete the task. In this work we will focus on two kinds o f environments: (nearly) reversible environments and environments that can be re set. Alice will "propose" the task by doing a sequence of actions and then Bob m ust undo or repeat them, respectively. Via an appropriate reward structure, Ali ce and Bob automatically generate a curriculum of exploration, enabling unsuperv ised training of the agent. When Bob is deployed on an RL task within the enviro nment, this unsupervised training reduces the number of supervised episodes need ed to learn, and in some cases converges to a higher reward.
****

On the Construction and Evaluation of Color Invariant Networks
Konrad Groh
This is an empirical paper which constructs color invariant networks and evaluat es their performances on a realistic data set. The paper studies the simplest po ssible case of color invariance: invariance under pixel-wise permutation of the color channels. Thus the network is aware not of the specific color object, but its colorfulness. The data set introduced in the paper consists of images showin g crashed cars from which ten classes were extracted. An additional annotation w as done which labeled whether the car shown was red or non-red. The networks we re evaluated by their performance on the classification task. With the color ann otation we altered the color ratios in the training data and analyzed the gener alization capabilities of the networks on the unaltered test data. We further sp lit the test data in red and non-red cars and did a similar evaluation. It is sh own in the paper that an pixel-wise ordering of the rgb-values of the images per forms better or at least similarly for small deviations from the true color rati os. The limits of these networks are also discussed.
****

Time Limits in Reinforcement Learning
Fabio Pardo,Arash Tavakoli,Vitaly Levdik,Petar Kormushev
In reinforcement learning, it is common to let an agent interact with its enviro nment for a fixed amount of time before resetting the environment and repeating the process in a series of episodes. The task that the agent has to learn can ei ther be to maximize its performance over (i) that fixed amount of time, or (ii) an indefinite period where the time limit is only used during training. In this paper, we investigate theoretically how time limits could effectively be handled in each of the two cases. In the first one, we argue that the terminations due to time limits are in fact part of the environment, and propose to include a not ion of the remaining time as part of the agent's input. In the second case, the

time limits are not part of the environment and are only used to facilitate lear ning. We argue that such terminations should not be treated as environmental one s and propose a method, specific to value-based algorithms, that incorporates th is insight by continuing to bootstrap at the end of each partial episode. To ill ustrate the significance of our proposals, we perform several experiments on a r ange of environments from simple few-state transition graphs to complex control tasks, including novel and standard benchmark domains. Our results show that the proposed methods improve the performance and stability of existing reinforcemen t learning algorithms.
****

LSH-SAMPLING BREAKS THE COMPUTATIONAL CHICKEN-AND-EGG LOOP IN ADAPTIVE STOCHASTI C GRADIENT ESTIMATION

Beidi Chen,Yingchen Xu,Anshumali Shrivastava

Stochastic Gradient Descent or SGD is the most popular optimization algorithm fo r large-scale problems. SGD estimates the gradient by uniform sampling with samp le size one. There have been several other works that suggest faster epoch wise convergence by using weighted non-uniform sampling for better gradient estimates . Unfortunately, the per-iteration cost of maintaining this adaptive distributio n for gradient estimation is more than calculating the full gradient. As a resul t, the false impression of faster convergence in iterations leads to slower conv ergence in time, which we call a chicken-and-egg loop. In this paper, we break t his barrier by providing the first demonstration of a sampling scheme, which lea ds to superior gradient estimation, while keeping the sampling cost per iteratio n similar to that of the uniform sampling. Such an algorithm is possible due to the sampling view of Locality Sensitive Hashing (LSH), which came to light recen tly. As a consequence of superior and fast estimation, we reduce the running tim e of all existing gradient descent algorithms. We demonstrate the benefits of ou r proposal on both SGD and AdaGrad.
****

Spontaneous Symmetry Breaking in Deep Neural Networks

Ricky Fok,Aijun An,Xiaogang Wang

We propose a framework to understand the unprecedented performance and robustnes s of deep neural networks using field theory. Correlations between the weights w ithin the same layer can be described by symmetries in that layer, and networks generalize better if such symmetries are broken to reduce the redundancies of th e weights. Using a two parameter field theory, we find that the network can brea k such symmetries itself towards the end of training in a process commonly known in physics as spontaneous symmetry breaking. This corresponds to a network gene ralizing itself without any user input layers to break the symmetry, but by comm unication with adjacent layers. In the layer decoupling limit applicable to resi dual networks (He et al., 2015), we show that the remnant symmetries that surviv e the non-linear layers are spontaneously broken based on empirical results. The Lagrangian for the non-linear and weight layers together has striking similarit ies with the one in quantum field theory of a scalar. Using results from quantum field theory we show that our framework is able to explain many experimentally observed phenomena, such as training on random labels with zero error (Zhang et al., 2017), the information bottleneck and the phase transition out of it (Shwar tz-Ziv & Tishby, 2017), shattered gradients (Balduzzi et al., 2017), and many mo re.
****

GeoSeq2Seq: Information Geometric Sequence-to-Sequence Networks

Alessandro Bay,Biswa Sengupta

The Fisher information metric is an important foundation of information geometry , wherein it allows us to approximate the local geometry of a probability distri bution. Recurrent neural networks such as the Sequence-to-Sequence (Seq2Seq) net works that have lately been used to yield state-of-the-art performance on speech translation or image captioning have so far ignored the geometry of the latent embedding, that they iteratively learn. We propose the information geometric Seq 2Seq (GeoSeq2Seq) network which abridges the gap between deep recurrent neural n etworks and information geometry. Specifically, the latent embedding offered by

a recurrent network is encoded as a Fisher kernel of a parametric Gaussian Mixture Model, a formalism common in computer vision. We utilise such a network to predict the shortest routes between two nodes of a graph by learning the adjacency matrix using the GeoSeq2Seq formalism; our results show that for such a problem the probabilistic representation of the latent embedding supersedes the non-probabilistic embedding by 10-15\%.
****
Learning Priors for Adversarial Autoencoders
Hui-Po Wang,Wei-Jan Ko,Wen-Hsiao Peng
Most deep latent factor models choose simple priors for simplicity, tractability or not knowing what prior to use. Recent studies show that the choice of the prior may have a profound effect on the expressiveness of the model, especially when its generative network has limited capacity. In this paper, we propose to learn a proper prior from data for adversarial autoencoders (AAEs). We introduce the notion of code generators to transform manually selected simple priors into ones that can better characterize the data distribution. Experimental results show that the proposed model can generate better image quality and learn better disentangled representations than AAEs in both supervised and unsupervised settings. Lastly, we present its ability to do cross-domain translation in a text-to-image synthesis task.
****
Large scale distributed neural network training through online distillation
Rohan Anil,Gabriel Pereyra,Alexandre Passos,Robert Ormandi,George E. Dahl,Geoffrey E. Hinton
Techniques such as ensembling and distillation promise model quality improvements when paired with almost any base model. However, due to increased test-time cost (for ensembles) and increased complexity of the training pipeline (for distillation), these techniques are challenging to use in industrial settings. In this paper we explore a variant of distillation which is relatively straightforward to use as it does not require a complicated multi-stage setup or many new hyperparameters. Our first claim is that online distillation enables us to use extra parallelism to fit very large datasets about twice as fast. Crucially, we can still speed up training even after we have already reached the point at which additional parallelism provides no benefit for synchronous or asynchronous stochastic gradient descent. Two neural networks trained on disjoint subsets of the data can share knowledge by encouraging each model to agree with the predictions the other model would have made. These predictions can come from a stale version of the other model so they can be safely computed using weights that only rarely get transmitted. Our second claim is that online distillation is a cost-effective way to make the exact predictions of a model dramatically more reproducible. We support our claims using experiments on the Criteo Display Ad Challenge dataset, ImageNet, and the largest to-date dataset used for neural language modeling, containing 6×10^{11} tokens and based on the Common Crawl repository of web data.
****
Lifelong Learning by Adjusting Priors
Ron Amit,Ron Meir
In representational lifelong learning an agent aims to continually learn to solve novel tasks while updating its representation in light of previous tasks. Under the assumption that future tasks are related to previous tasks, representations should be learned in such a way that they capture the common structure across learned tasks, while allowing the learner sufficient flexibility to adapt to novel aspects of a new task. We develop a framework for lifelong learning in deep neural networks that is based on generalization bounds, developed within the PAC-Bayes framework. Learning takes place through the construction of a distribution over networks based on the tasks seen so far, and its utilization for learning a new task. Thus, prior knowledge is incorporated through setting a history-dependent prior for novel tasks. We develop a gradient-based algorithm implementing these ideas, based on minimizing an objective function motivated by generalizati

on bounds, and demonstrate its effectiveness through numerical examples.
****
Variance Regularized Counterfactual Risk Minimization via Variational Divergence Minimization
Hang Wu
Off-policy learning, the task of evaluating and improving policies using historic data collected from a logging policy, is important because on-policy evaluation is usually expensive and has adverse impacts. One of the major challenge of off-policy learning is to derive counterfactual estimators that also has low variance and thus low generalization error.
In this work, inspired by learning bounds for importance sampling problems, we present a new counterfactual learning principle for off-policy learning with bandit feedbacks.Our method regularizes the generalization error by minimizing the distribution divergence between the logging policy and the new policy, and removes the need for iterating through all training samples to compute sample variance regularization in prior work. With neural network policies, our end-to-end training algorithms using variational divergence minimization showed significant improvement over conventional baseline algorithms and is also consistent with our theoretical results.
****
Universal Agent for Disentangling Environments and Tasks
Jiayuan Mao,Honghua Dong,Joseph J. Lim
Recent state-of-the-art reinforcement learning algorithms are trained under the goal of excelling in one specific task. Hence, both environment and task specific knowledge are entangled into one framework. However, there are often scenarios where the environment (e.g. the physical world) is fixed while only the target task changes. Hence, borrowing the idea from hierarchical reinforcement learning, we propose a framework that disentangles task and environment specific knowledge by separating them into two units. The environment-specific unit handles how to move from one state to the target state; and the task-specific unit plans for the next target state given a specific task. The extensive results in simulators indicate that our method can efficiently separate and learn two independent units, and also adapt to a new task more efficiently than the state-of-the-art methods.
****
The Set Autoencoder: Unsupervised Representation Learning for Sets
Malte Probst
We propose the set autoencoder, a model for unsupervised representation learning for sets of elements. It is closely related to sequence-to-sequence models, which learn fixed-sized latent representations for sequences, and have been applied to a number of challenging supervised sequence tasks such as machine translation, as well as unsupervised representation learning for sequences.
In contrast to sequences, sets are permutation invariant. The proposed set autoencoder considers this fact, both with respect to the input as well as the output of the model. On the input side, we adapt a recently-introduced recurrent neural architecture using a content-based attention mechanism. On the output side, we use a stable marriage algorithm to align predictions to labels in the learning phase.
We train the model on synthetic data sets of point clouds and show that the learned representations change smoothly with translations in the inputs, preserve distances in the inputs, and that the set size is represented directly. We apply the model to supervised tasks on the point clouds using the fixed-size latent representation. For a number of difficult classification problems, the results are better than those of a model that does not consider the permutation invariance. Especially for small training sets, the set-aware model benefits from unsupervised pretraining.
****
FusionNet: Fusing via Fully-aware Attention with Application to Machine Comprehension
Hsin-Yuan Huang,Chenguang Zhu,Yelong Shen,Weizhu Chen

This paper introduces a new neural structure called FusionNet, which extends exi
sting attention approaches from three perspectives. First, it puts forward a nov
el concept of "History of Word" to characterize attention information from the l
owest word-level embedding up to the highest semantic-level representation. Seco
nd, it identifies an attention scoring function that better utilizes the "histor
y of word" concept. Third, it proposes a fully-aware multi-level attention mecha
nism to capture the complete information in one text (such as a question) and ex
ploit it in its counterpart (such as context or passage) layer by layer. We appl
y FusionNet to the Stanford Question Answering Dataset (SQuAD) and it achieves t
he first position for both single and ensemble model on the official SQuAD leade
rboard at the time of writing (Oct. 4th, 2017). Meanwhile, we verify the general
ization of FusionNet with two adversarial SQuAD datasets and it sets up the new
state-of-the-art on both datasets: on AddSent, FusionNet increases the best F1 m
etric from 46.6% to 51.4%; on AddOneSent, FusionNet boosts the best F1 metric fr
om 56.0% to 60.7%.
****
Benefits of Depth for Long-Term Memory of Recurrent Networks
Yoav Levine,Or Sharir,Amnon Shashua
The key attribute that drives the unprecedented success of modern Recurrent Neur
al Networks (RNNs) on learning tasks which involve sequential data, is their eve
r-improving ability to model intricate long-term temporal dependencies. However,
 a well established measure of RNNs' long-term memory capacity is lacking, and t
hus formal understanding of their ability to correlate data throughout time is l
imited. Though depth efficiency in convolutional networks is well established by
 now, it does not suffice in order to account for the success of deep RNNs on in
puts of varying lengths, and the need to address their 'time-series expressive p
ower' arises. In this paper, we analyze the effect of depth on the ability of re
current networks to express correlations ranging over long time-scales. To meet
the above need, we introduce a measure of the information flow across time that
can be supported by the network, referred to as the Start-End separation rank. E
ssentially, this measure reflects the distance of the function realized by the r
ecurrent network from a function that models no interaction whatsoever between t
he beginning and end of the input sequence. We prove that deep recurrent network
s support Start-End separation ranks which are exponentially higher than those s
upported by their shallow counterparts. Moreover, we show that the ability of de
ep recurrent networks to correlate different parts of the input sequence increas
es exponentially as the input sequence extends, while that of vanilla shallow re
current networks does not adapt to the sequence length at all. Thus, we establis
h that depth brings forth an overwhelming advantage in the ability of recurrent
networks to model long-term dependencies, and provide an exemplar of quantifying
 this key attribute which may be readily extended to other RNN architectures of
interest, e.g. variants of LSTM networks. We obtain our results by considering a
 class of recurrent networks referred to as Recurrent Arithmetic Circuits (RACs)
, which merge the hidden state with the input via the Multiplicative Integration
 operation.
****
Data Augmentation Generative Adversarial Networks
Anthreas Antoniou,Amos Storkey,Harrison Edwards
Effective training of neural networks requires much data. In the low-data regime
,
parameters are underdetermined, and learnt networks generalise poorly. Data
Augmentation (Krizhevsky et al., 2012) alleviates this by using existing data
more effectively. However standard data augmentation produces only limited
plausible alternative data. Given there is potential to generate a much broader
set
of augmentations, we design and train a generative model to do data augmentation
.
The model, based on image conditional Generative Adversarial Networks, takes
data from a source domain and learns to take any data item and generalise it
to generate other within-class data items. As this generative process does not

depend on the classes themselves, it can be applied to novel unseen classes of data.
We show that a Data Augmentation Generative Adversarial Network (DAGAN) augments standard vanilla classifiers well. We also show a DAGAN can enhance few-shot learning systems such as Matching Networks. We demonstrate these approaches on Omniglot, on EMNIST having learnt the DAGAN on Omniglot, and VGG-Face data. In our experiments we can see over 13% increase in accuracy in the low-data regime experiments in Omniglot (from 69% to 82%), EMNIST (73.9% to 76%) and VGG-Face (4.5% to 12%); in Matching Networks for Omniglot we observe an increase of 0.5% (from 96.9% to 97.4%) and an increase of 1.8% in EMNIST (from 59.5% to 61.3%).
****
Regret Minimization for Partially Observable Deep Reinforcement Learning
Peter H. Jin,Sergey Levine,Kurt Keutzer
Deep reinforcement learning algorithms that estimate state and state-action value functions have been shown to be effective in a variety of challenging domains, including learning control strategies from raw image pixels. However, algorithms that estimate state and state-action value functions typically assume a fully observed state and must compensate for partial or non-Markovian observations by using finite-length frame-history observations or recurrent networks. In this work, we propose a new deep reinforcement learning algorithm based on counterfactual regret minimization that iteratively updates an approximation to a cumulative clipped advantage function and is robust to partially observed state. We demonstrate that on several partially observed reinforcement learning tasks, this new class of algorithms can substantially outperform strong baseline methods: on Pong with single-frame observations, and on the challenging Doom (ViZDoom) and Minecraft (Malmö) first-person navigation benchmarks.
****
Global Optimality Conditions for Deep Neural Networks
Chulhee Yun,Suvrit Sra,Ali Jadbabaie
We study the error landscape of deep linear and nonlinear neural networks with the squared error loss. Minimizing the loss of a deep linear neural network is a nonconvex problem, and despite recent progress, our understanding of this loss surface is still incomplete. For deep linear networks, we present necessary and sufficient conditions for a critical point of the risk function to be a global minimum. Surprisingly, our conditions provide an efficiently checkable test for global optimality, while such tests are typically intractable in nonconvex optimization. We further extend these results to deep nonlinear neural networks and prove similar sufficient conditions for global optimality, albeit in a more limited function space setting.
****
Guide Actor-Critic for Continuous Control
Voot Tangkaratt,Abbas Abdolmaleki,Masashi Sugiyama
Actor-critic methods solve reinforcement learning problems by updating a parameterized policy known as an actor in a direction that increases an estimate of the expected return known as a critic. However, existing actor-critic methods only use values or gradients of the critic to update the policy parameter. In this paper, we propose a novel actor-critic method called the guide actor-critic (GAC). GAC firstly learns a guide actor that locally maximizes the critic and then it updates the policy parameter based on the guide actor by supervised learning. Our main theoretical contributions are two folds. First, we show that GAC updates the guide actor by performing second-order optimization in the action space where the curvature matrix is based on the Hessians of the critic. Second, we show that the deterministic policy gradient method is a special case of GAC when the Hessians are ignored. Through experiments, we show that our method is a promising reinforcement learning method for continuous controls.

****
A Self-Organizing Memory Network
Callie Federer,Joel Zylberberg

Working memory requires information about external stimuli to be represented in the brain even after those stimuli go away. This information is encoded in the activities of neurons, and neural activities change over timescales of tens of milliseconds. Information in working memory, however, is retained for tens of seconds, suggesting the question of how time-varying neural activities maintain stable representations. Prior work shows that, if the neural dynamics are in the ` null space' of the representation - so that changes to neural activity do not affect the downstream read-out of stimulus information - then information can be retained for periods much longer than the time-scale of individual-neuronal activities. The prior work, however, requires precisely constructed synaptic connectivity matrices, without explaining how this would arise in a biological neural network. To identify mechanisms through which biological networks can self-organize to learn memory function, we derived biologically plausible synaptic plasticity rules that dynamically modify the connectivity matrix to enable information storing. Networks implementing this plasticity rule can successfully learn to form memory representations even if only 10% of the synapses are plastic, they are robust to synaptic noise, and they can represent information about multiple stimuli.
****

Kronecker-factored Curvature Approximations for Recurrent Neural Networks
James Martens,Jimmy Ba,Matt Johnson
Kronecker-factor Approximate Curvature (Martens & Grosse, 2015) (K-FAC) is a 2nd-order optimization method which has been shown to give state-of-the-art performance on large-scale neural network optimization tasks (Ba et al., 2017). It is based on an approximation to the Fisher information matrix (FIM) that makes assumptions about the particular structure of the network and the way it is parameterized. The original K-FAC method was applicable only to fully-connected networks, although it has been recently extended by Grosse & Martens (2016) to handle convolutional networks as well. In this work we extend the method to handle RNNs by introducing a novel approximation to the FIM for RNNs. This approximation works by modelling the covariance structure between the gradient contributions at different time-steps using a chain-structured linear Gaussian graphical model, summing the various cross-covariances, and computing the inverse in closed form. We demonstrate in experiments that our method significantly outperforms general purpose state-of-the-art optimizers like SGD with momentum and Adam on several challenging RNN training tasks.
****

Now I Remember! Episodic Memory For Reinforcement Learning
Ricky Loynd,Matthew Hausknecht,Lihong Li,Li Deng
Humans rely on episodic memory constantly, in remembering the name of someone they met 10 minutes ago, the plot of a movie as it unfolds, or where they parked the car. Endowing reinforcement learning agents with episodic memory is a key step on the path toward replicating human-like general intelligence. We analyze why standard RL agents lack episodic memory today, and why existing RL tasks don't require it. We design a new form of external memory called Masked Experience Memory, or MEM, modeled after key features of human episodic memory. To evaluate episodic memory we define an RL task based on the common children's game of Concentration. We find that a MEM RL agent leverages episodic memory effectively to master Concentration, unlike the baseline agents we tested.
****

Learning a neural response metric for retinal prosthesis
Nishal P Shah,Sasidhar Madugula,EJ Chichilnisky,Yoram Singer,Jonathon Shlens
Retinal prostheses for treating incurable blindness are designed to electrically stimulate surviving retinal neurons, causing them to send artificial visual signals to the brain. However, electrical stimulation generally cannot precisely reproduce normal patterns of neural activity in the retina. Therefore, an electrical stimulus must be selected that produces a neural response as close as possible to the desired response. This requires a technique for computing a distance between the desired response and the achievable response that is meaningful in terms of the visual signal being conveyed. Here we propose a method to learn such

a metric on neural responses, directly from recorded light responses of a population of retinal ganglion cells (RGCs) in the primate retina. The learned metric produces a measure of similarity of RGC population responses that accurately reflects the similarity of the visual input. Using data from electrical stimulation experiments, we demonstrate that this metric may improve the performance of a prosthesis.
****

Empirical Analysis of the Hessian of Over-Parametrized Neural Networks
Levent Sagun,Utku Evci,V. Ugur Guney,Yann Dauphin,Leon Bottou
We study the properties of common loss surfaces through their Hessian matrix. In particular, in the context of deep learning, we empirically show that the spectrum of the Hessian is composed of two parts: (1) the bulk centered near zero, (2) and outliers away from the bulk. We present numerical evidence and mathematical justifications to the following conjectures laid out by Sagun et. al. (2016): Fixing data, increasing the number of parameters merely scales the bulk of the spectrum; fixing the dimension and changing the data (for instance adding more clusters or making the data less separable) only affects the outliers. We believe that our observations have striking implications for non-convex optimization in high dimensions. First, the *flatness* of such landscapes (which can be measured by the singularity of the Hessian) implies that classical notions of basins of attraction may be quite misleading. And that the discussion of wide/narrow basins may be in need of a new perspective around over-parametrization and redundancy that are able to create *large* connected components at the bottom of the landscape. Second, the dependence of a small number of large eigenvalues to the data distribution can be linked to the spectrum of the covariance matrix of gradients of model outputs. With this in mind, we may reevaluate the connections within the data-architecture-algorithm framework of a model, hoping that it would shed light on the geometry of high-dimensional and non-convex spaces in modern applications. In particular, we present a case that links the two observations: small and large batch gradient descent appear to converge to different basins of attraction but we show that they are in fact connected through their flat region and so belong to the same basin.
****

Quadrature-based features for kernel approximation
Marina Munkhoeva,Yermek Kapushev,Evgeny Burnaev,Ivan Oseledets
We consider the problem of improving kernel approximation via feature maps. These maps arise as Monte Carlo approximation to integral representations of kernel functions and scale up kernel methods for larger datasets. We propose to use more efficient numerical integration technique to obtain better estimates of the integrals compared to the state-of-the-art methods. Our approach allows to use information about the integrand to enhance approximation and facilitates fast computations. We derive the convergence behavior and conduct an extensive empirical study that supports our hypothesis.
****

Certified Defenses against Adversarial Examples
Aditi Raghunathan,Jacob Steinhardt,Percy Liang
While neural networks have achieved high accuracy on standard image classification benchmarks, their accuracy drops to nearly zero in the presence of small adversarial perturbations to test inputs. Defenses based on regularization and adversarial training have been proposed, but often followed by new, stronger attacks that defeat these defenses. Can we somehow end this arms race? In this work, we study this problem for neural networks with one hidden layer. We first propose a method based on a semidefinite relaxation that outputs a certificate that for a given network and test input, no attack can force the error to exceed a certain value. Second, as this certificate is differentiable, we jointly optimize it with the network parameters, providing an adaptive regularizer that encourages robustness against all attacks. On MNIST, our approach produces a network and a certificate that no that perturbs each pixel by at most $\epsilon = 0.1$ can cause more than 35% test error.

****
DDRprog: A CLEVR Differentiable Dynamic Reasoning Programmer
Joseph Suarez,Justin Johnson,L. Fei-Fei
We present a generic dynamic architecture that employs a problem specific differentiable forking mechanism to leverage discrete logical information about the problem data structure. We adapt and apply our model to CLEVR Visual Question Answering, giving rise to the DDRprog architecture; compared to previous approaches, our model achieves higher accuracy in half as many epochs with five times fewer learnable parameters. Our model directly models underlying question logic using a recurrent controller that jointly predicts and executes functional neural modules; it explicitly forks subprocesses to handle logical branching. While FiLM and other competitive models are static architectures with less supervision, we argue that inclusion of program labels enables learning of higher level logical operations -- our architecture achieves particularly high performance on questions requiring counting and integer comparison. We further demonstrate the generality of our approach though DDRstack -- an application of our method to reverse Polish notation expression evaluation in which the inclusion of a stack assumption allows our approach to generalize to long expressions, significantly outperforming an LSTM with ten times as many learnable parameters.
****
Activation Maximization Generative Adversarial Nets
Zhiming Zhou,Han Cai,Shu Rong,Yuxuan Song,Kan Ren,Weinan Zhang,Jun Wang,Yong Yu
Class labels have been empirically shown useful in improving the sample quality of generative adversarial nets (GANs). In this paper, we mathematically study the properties of the current variants of GANs that make use of class label information. With class aware gradient and cross-entropy decomposition, we reveal how class labels and associated losses influence GAN's training. Based on that, we propose Activation Maximization Generative Adversarial Networks (AM-GAN) as an advanced solution. Comprehensive experiments have been conducted to validate our analysis and evaluate the effectiveness of our solution, where AM-GAN outperforms other strong baselines and achieves state-of-the-art Inception Score (8.91) on CIFAR-10. In addition, we demonstrate that, with the Inception ImageNet classifier, Inception Score mainly tracks the diversity of the generator, and there is, however, no reliable evidence that it can reflect the true sample quality. We thus propose a new metric, called AM Score, to provide more accurate estimation on the sample quality. Our proposed model also outperforms the baseline methods in the new metric.
****
Reward Design in Cooperative Multi-agent Reinforcement Learning for Packet Routing
Hangyu Mao,Zhibo Gong,Zhen Xiao
In cooperative multi-agent reinforcement learning (MARL), how to design a suitable reward signal to accelerate learning and stabilize convergence is a critical problem. The global reward signal assigns the same global reward to all agents without distinguishing their contributions, while the local reward signal provides different local rewards to each agent based solely on individual behavior. Both of the two reward assignment approaches have some shortcomings: the former might encourage lazy agents, while the latter might produce selfish agents.

In this paper, we study reward design problem in cooperative MARL based on packet routing environments. Firstly, we show that the above two reward signals are prone to produce suboptimal policies. Then, inspired by some observations and considerations, we design some mixed reward signals, which are off-the-shelf to learn better policies. Finally, we turn the mixed reward signals into the adaptive counterparts, which achieve best results in our experiments. Other reward signals are also discussed in this paper. As reward design is a very fundamental problem in RL and especially in MARL, we hope that MARL researchers can rethink the rewards used in their systems.
****
Disentangled activations in deep networks

Mikael Kågebäck,Olof Mogren
Deep neural networks have been tremendously successful in a number of tasks.
One of the main reasons for this is their capability to automatically
learn representations of data in levels of abstraction,
increasingly disentangling the data as the internal transformations are applied.
In this paper we propose a novel regularization method that penalize covariance between dimensions of the hidden layers in a network, something that benefits the disentanglement.
This makes the network learn nonlinear representations that are linearly uncorrelated, yet allows the model to obtain good results on a number of tasks, as demonstrated by our experimental evaluation.
The proposed technique can be used to find the dimensionality of the underlying data, because it effectively disables dimensions that aren't needed.
Our approach is simple and computationally cheap, as it can be applied as a regularizer to any gradient-based learning model.
****
Evidence Aggregation for Answer Re-Ranking in Open-Domain Question Answering
Shuohang Wang,Mo Yu,Jing Jiang,Wei Zhang,Xiaoxiao Guo,Shiyu Chang,Zhiguo Wang,Tim Klinger,Gerald Tesauro,Murray Campbell
Very recently, it comes to be a popular approach for answering open-domain questions by first searching question-related passages, then applying reading comprehension models to extract answers. Existing works usually extract answers from single passages independently, thus not fully make use of the multiple searched passages, especially for the some questions requiring several evidences, which can appear in different passages, to be answered. The above observations raise the problem of evidence aggregation from multiple passages. In this paper, we deal with this problem as answer re-ranking. Specifically, based on the answer candidates generated from the existing state-of-the-art QA model, we propose two different re-ranking methods, strength-based and coverage-based re-rankers, which make use of the aggregated evidences from different passages to help entail the ground-truth answer for the question. Our model achieved state-of-the-arts on three public open-domain QA datasets, Quasar-T, SearchQA and the open-domain version of TriviaQA, with about 8\% improvement on the former two datasets.
****
Few-Shot Learning with Graph Neural Networks
Victor Garcia Satorras,Joan Bruna Estrach
We propose to study the problem of few-shot learning with the prism of inference on a partially observed graphical model, constructed from a collection of input images whose label can be either observed or not. By assimilating generic message-passing inference algorithms with their neural-network counterparts, we define a graph neural network architecture that generalizes several of the recently proposed few-shot learning models. Besides providing improved numerical performance, our framework is easily extended to variants of few-shot learning, such as semi-supervised or active learning, demonstrating the ability of graph-based models to operate well on ‘relational’ tasks.
****
Softmax Q-Distribution Estimation for Structured Prediction: A Theoretical Interpretation for RAML
Xuezhe Ma,Pengcheng Yin,Jingzhou Liu,Graham Neubig,Eduard Hovy
Reward augmented maximum likelihood (RAML), a simple and effective learning framework to directly optimize towards the reward function in structured prediction tasks, has led to a number of impressive empirical successes. RAML incorporates task-specific reward by performing maximum-likelihood updates on candidate outputs sampled according to an exponentiated payoff distribution, which gives higher probabilities to candidates that are close to the reference output. While RAML is notable for its simplicity, efficiency, and its impressive empirical successes, the theoretical properties of RAML, especially the behavior of the exponentiated payoff distribution, has not been examined thoroughly. In this work, we introduce softmax Q-distribution estimation, a novel theoretical interpretation of RAML, which reveals the relation between RAML and Bayesian decision theory. The s

oftmax Q-distribution can be regarded as a smooth approximation of the Bayes decision boundary, and the Bayes decision rule is achieved by decoding with this Q-distribution. We further show that RAML is equivalent to approximately estimating the softmax Q-distribution, with the temperature τ controlling approximation error. We perform two experiments, one on synthetic data of multi-class classification and one on real data of image captioning, to demonstrate the relationship between RAML and the proposed softmax Q-distribution estimation, verifying our theoretical analysis. Additional experiments on three structured prediction tasks with rewards defined on sequential (named entity recognition), tree-based (dependency parsing) and irregular (machine translation) structures show notable improvements over maximum likelihood baselines.
****
Action-dependent Control Variates for Policy Optimization via Stein Identity
Hao Liu*,Yihao Feng*,Yi Mao,Dengyong Zhou,Jian Peng,Qiang Liu
Policy gradient methods have achieved remarkable successes in solving challenging reinforcement learning problems. However, it still often suffers from the large variance issue on policy gradient estimation, which leads to poor sample efficiency during training. In this work, we propose a control variate method to effectively reduce variance for policy gradient methods. Motivated by the Stein's identity, our method extends the previous control variate methods used in REINFORCE and advantage actor-critic by introducing more flexible and general action-dependent baseline functions. Empirical studies show that our method essentially improves the sample efficiency of the state-of-the-art policy gradient approaches.

****
Distributional Adversarial Networks
Chengtao Li,David Alvarez-Melis,Keyulu Xu,Stefanie Jegelka,Suvrit Sra
In most current formulations of adversarial training, the discriminators can be expressed as single-input operators, that is, the mapping they define is separable over observations. In this work, we argue that this property might help explain the infamous mode collapse phenomenon in adversarially-trained generative models. Inspired by discrepancy measures and two-sample tests between probability distributions, we propose distributional adversaries that operate on samples, i.e., on sets of multiple points drawn from a distribution, rather than on single observations. We show how they can be easily implemented on top of existing models. Various experimental results show that generators trained in combination with our distributional adversaries are much more stable and are remarkably less prone to mode collapse than traditional models trained with observation-wise prediction discriminators. In addition, the application of our framework to domain adaptation results in strong improvement over recent state-of-the-art.
****
Decision Boundary Analysis of Adversarial Examples
Warren He,Bo Li,Dawn Song
Deep neural networks (DNNs) are vulnerable to adversarial examples, which are carefully crafted instances aiming to cause prediction errors for DNNs. Recent research on adversarial examples has examined local neighborhoods in the input space of DNN models. However, previous work has limited what regions to consider, focusing either on low-dimensional subspaces or small balls. In this paper, we argue that information from larger neighborhoods, such as from more directions and from greater distances, will better characterize the relationship between adversarial examples and the DNN models. First, we introduce an attack, OPTMARGIN, which generates adversarial examples robust to small perturbations. These examples successfully evade a defense that only considers a small ball around an input instance. Second, we analyze a larger neighborhood around input instances by looking at properties of surrounding decision boundaries, namely the distances to the boundaries and the adjacent classes. We find that the boundaries around these adversarial examples do not resemble the boundaries around benign examples. Finally, we show that, under scrutiny of the surrounding decision boundaries, our OPTMARGIN examples do not convincingly mimic benign examples. Although our experiments are limited to a few specific attacks, we hope these findings will motivate

new, more evasive attacks and ultimately, effective defenses.
****
Relevance of Unsupervised Metrics in Task-Oriented Dialogue for Evaluating Natural Language Generation
Shikhar Sharma,Layla El Asri,Hannes Schulz,Jeremie Zumer
Automated metrics such as BLEU are widely used in the machine translation literature. They have also been used recently in the dialogue community for evaluating dialogue response generation. However, previous work in dialogue response generation has shown that these metrics do not correlate strongly with human judgment in the non task-oriented dialogue setting. Task-oriented dialogue responses are expressed on narrower domains and exhibit lower diversity. It is thus reasonable to think that these automated metrics would correlate well with human judgment in the task-oriented setting where the generation task consists of translating dialogue acts into a sentence. We conduct an empirical study to confirm whether this is the case. Our findings indicate that these automated metrics have stronger correlation with human judgments in the task-oriented setting compared to what has been observed in the non task-oriented setting. We also observe that these metrics correlate even better for datasets which provide multiple ground truth reference sentences. In addition, we show that some of the currently available corpora for task-oriented language generation can be solved with simple models and advocate for more challenging datasets.
****
Memory-based Parameter Adaptation
Pablo Sprechmann,Siddhant M. Jayakumar,Jack W. Rae,Alexander Pritzel,Adria Puigdomenech Badia,Benigno Uria,Oriol Vinyals,Demis Hassabis,Razvan Pascanu,Charles Blundell
Deep neural networks have excelled on a wide range of problems, from vision to language and game playing. Neural networks very gradually incorporate information into weights as they process data, requiring very low learning rates. If the training distribution shifts, the network is slow to adapt, and when it does adapt, it typically performs badly on the training distribution before the shift. Our method, Memory-based Parameter Adaptation, stores examples in memory and then uses a context-based lookup to directly modify the weights of a neural network. Much higher learning rates can be used for this local adaptation, reneging the need for many iterations over similar data before good predictions can be made. As our method is memory-based, it alleviates several shortcomings of neural networks, such as catastrophic forgetting, fast, stable acquisition of new knowledge, learning with an imbalanced class labels, and fast learning during evaluation. We demonstrate this on a range of supervised tasks: large-scale image classification and language modelling.
****
Latent Topic Conversational Models
Tsung-Hsien Wen,Minh-Thang Luong
Despite much success in many large-scale language tasks, sequence-to-sequence (seq2seq) models have not been an ideal choice for conversational modeling as they tend to generate generic and repetitive responses. In this paper, we propose a Latent Topic Conversational Model (LTCM) that augments the seq2seq model with a neural topic component to better model human-human conversations. The neural topic component encodes information from the source sentence to build a global “topic” distribution over words, which is then consulted by the seq2seq model to improve generation at each time step. The experimental results show that the proposed LTCM can generate more diverse and interesting responses by sampling from its learnt latent representations. In a subjective human evaluation, the judges also confirm that LTCM is the preferred option comparing to competitive baseline models.

****
FearNet: Brain-Inspired Model for Incremental Learning
Ronald Kemker,Christopher Kanan
Incremental class learning involves sequentially learning classes in bursts of e

xamples from the same class. This violates the assumptions that underlie method
s for training standard deep neural networks, and will cause them to suffer from
 catastrophic forgetting. Arguably, the best method for incremental class learni
ng is iCaRL, but it requires storing training examples for each class, making i
t challenging to scale. Here, we propose FearNet for incremental class learning.
 FearNet is a generative model that does not store previous examples, making it
memory efficient. FearNet uses a brain-inspired dual-memory system in which new
memories are consolidated from a network for recent memories inspired by the mam
malian hippocampal complex to a network for long-term storage inspired by medial
 prefrontal cortex. Memory consolidation is inspired by mechanisms that occur du
ring sleep. FearNet also uses a module inspired by the basolateral amygdala for
determining which memory system to use for recall. FearNet achieves state-of-th
e-art performance at incremental class learning on image (CIFAR-100, CUB-200) an
d audio classification (AudioSet) benchmarks.

****

 Neural Map: Structured Memory for Deep Reinforcement Learning
Emilio Parisotto,Ruslan Salakhutdinov
A critical component to enabling intelligent reasoning in partially observable e
nvironments is memory. Despite this importance, Deep Reinforcement Learning (DRL
) agents have so far used relatively simple memory architectures, with the main
methods to overcome partial observability being either a temporal convolution ov
er the past k frames or an LSTM layer. More recent work (Oh et al., 2016) has we
nt beyond these architectures by using memory networks which can allow more soph
isticated addressing schemes over the past k frames. But even these architecture
s are unsatisfactory due to the reason that they are limited to only remembering
 information from the last k frames. In this paper, we develop a memory system w
ith an adaptable write operator that is customized to the sorts of 3D environmen
ts that DRL agents typically interact with. This architecture, called the Neural
 Map, uses a spatially structured 2D memory image to learn to store arbitrary in
formation about the environment over long time lags. We demonstrate empirically
that the Neural Map surpasses previous DRL memories on a set of challenging 2D a
nd 3D maze environments and show that it is capable of generalizing to environme
nts that were not seen during training.

****

Understanding GANs: the LQG Setting
Soheil Feizi,Changho Suh,Fei Xia,David Tse
Generative Adversarial Networks (GANs) have become a popular method to learn a p
robability model from data. Many GAN architectures with different optimization m
etrics have been introduced recently. Instead of proposing yet another architect
ure, this paper aims to provide an understanding of some of the basic issues su
rrounding GANs. First, we propose a natural way of specifying the loss function
for GANs by drawing a connection with supervised learning. Second, we shed light
 on the statistical performance of GANs through the analysis of a simple LQG set
ting: the generator is linear, the loss function is quadratic and the data is dr
awn from a Gaussian distribution. We show that in this setting: 1) the optimal
GAN solution converges to population Principal Component Analysis (PCA) as the n
umber of training samples increases; 2) the number of samples required scales ex
ponentially with the dimension of the data; 3) the number of samples scales almo
st linearly if the discriminator is constrained to be quadratic. Moreover, under
 this quadratic constraint on the discriminator, the optimal finite-sample GAN p
erforms simply empirical PCA.

****

A Deep Learning Approach for Survival Clustering without End-of-life Signals
S Chandra Mouli,Bruno Ribeiro,Jennifer Neville
The goal of survival clustering is to map subjects (e.g., users in a social netw
ork, patients in a medical study) to K clusters ranging from low-risk to high-
risk. Existing survival methods assume the presence of clear \textit{end-of-life
} signals or introduce them artificially using a pre-defined timeout. In this pa
per, we forego this assumption and introduce a loss function that differentiates

between the empirical lifetime distributions of the clusters using a modified K
uiper statistic. We learn a deep neural network by optimizing this loss, that pe
rforms a soft clustering of users into survival groups. We apply our method to a
 social network dataset with over 1M subjects, and show significant improvement
in C-index compared to alternatives.
****
Iterative temporal differencing with fixed random feedback alignment support spi
ke-time dependent plasticity in vanilla backpropagation for deep learning
Aras Dargazany,Kunal Mankodiya
In vanilla backpropagation (VBP), activation function matters considerably in te
rms of non-linearity and differentiability.
Vanishing gradient has been an important problem related to the bad choice of ac
tivation function in deep learning (DL).
This work shows that a differentiable activation function is not necessary any m
ore for error backpropagation.
The derivative of the activation function can be replaced by an iterative tempor
al differencing (ITD) using fixed random feedback weight alignment (FBA).
Using FBA with ITD, we can transform the VBP into a more biologically plausible
approach for learning deep neural network architectures.
We don't claim that ITD works completely the same as the spike-time dependent pl
asticity (STDP) in our brain but this work can be a step toward the integration
of STDP-based error backpropagation in deep learning.
****
Combining Model-based and Model-free RL via Multi-step Control Variates
Tong Che,Yuchen Lu,George Tucker,Surya Bhupatiraju,Shane Gu,Sergey Levine,Yoshua
 Bengio
Model-free deep reinforcement learning algorithms are able to successfully solve
 a wide range of continuous control tasks, but typically require many on-policy
samples to achieve good performance. Model-based RL algorithms are sample-effici
ent on the other hand, while learning accurate global models of complex dynamic
environments has turned out to be tricky in practice, which leads to the unsatis
factory performance of the learned policies. In this work, we combine the sample
-efficiency of model-based algorithms and the accuracy of model-free algorithms.
 We leverage multi-step neural network based predictive models by embedding real
 trajectories into imaginary rollouts of the model, and use the imaginary cumula
tive rewards as control variates for model-free algorithms. In this way, we achi
eved the strengths of both sides and derived an estimator which is not only samp
le-efficient, but also unbiased and of very low variance. We present our evaluat
ion on the MuJoCo and OpenAI Gym benchmarks.
****
Reinforcement Learning from Imperfect Demonstrations
Yang Gao,Huazhe(Harry) Xu,Ji Lin,Fisher Yu,Sergey Levine,Trevor Darrell
Robust real-world learning should benefit from both demonstrations and interact
ion with the environment. Current approaches to learning from demonstration and
reward perform supervised learning on expert demonstration data and use reinforc
ement learning to further improve performance based on reward from the environm
ent. These tasks have divergent losses which are difficult to jointly optimize;
further, such methods can be very sensitive to noisy demonstrations. We propose
a unified reinforcement learning algorithm that effectively normalizes the Q-fun
ction, reducing the Q-values of actions unseen in the demonstration data. Our N
ormalized Actor-Critic (NAC) method can learn from demonstration data of arbitra
ry quality and also leverages rewards from an interactive environment. NAC lear
ns an initial policy network from demonstration and refines the policy in a real
 environment. Crucially, both learning from demonstration and interactive refine
ment use exactly the same objective, unlike prior approaches that combine distin
ct supervised and reinforcement losses. This makes NAC robust to suboptimal demo
nstration data, since the method is not forced to mimic all of the examples in t
he dataset. We show that our unified reinforcement learning algorithm can learn
robustly and outperform existing baselines when evaluated on several realistic
driving games.

****
Interpretable and Pedagogical Examples
Smitha Milli,Pieter Abbeel,Igor Mordatch
Teachers intentionally pick the most informative examples to show their students. However, if the teacher and student are neural networks, the examples that the teacher network learns to give, although effective at teaching the student, are typically uninterpretable. We show that training the student and teacher iteratively, rather than jointly, can produce interpretable teaching strategies. We evaluate interpretability by (1) measuring the similarity of the teacher's emergent strategies to intuitive strategies in each domain and (2) conducting human experiments to evaluate how effective the teacher's strategies are at teaching humans. We show that the teacher network learns to select or generate interpretable, pedagogical examples to teach rule-based, probabilistic, boolean, and hierarchical concepts.
****
THE EFFECTIVENESS OF A TWO-LAYER NEURAL NETWORK FOR RECOMMENDATIONS
Oleg Rybakov,Vijai Mohan,Avishkar Misra,Scott LeGrand,Rejith Joseph,Kiuk Chung,Siddharth Singh,Qian You,Eric Nalisnick,Leo Dirac,Runfei Luo
We present a personalized recommender system using neural network for recommending
products, such as eBooks, audio-books, Mobile Apps, Video and Music.
It produces recommendations based on customer’s implicit feedback history such as purchases, listens or watches. Our key contribution is to formulate recommendation
problem as a model that encodes historical behavior to predict the future behavior using soft data split, combining predictor and auto-encoder models. We introduce convolutional layer for learning the importance (time decay) of the purchases
depending on their purchase date and demonstrate that the shape of the time decay function can be well approximated by a parametrical function. We present offline experimental results showing that neural networks with two hidden layers can capture seasonality changes, and at the same time outperform other modeling techniques, including our recommender in production. Most importantly, we demonstrate that our model can be scaled to all digital categories, and we observe
significant improvements in an online A/B test. We also discuss key enhancements to the neural network model and describe our production pipeline. Finally
we open-sourced our deep learning library which supports multi-gpu model parallel
training. This is an important feature in building neural network based recommenders
with large dimensionality of input and output data.
****
APPLICATION OF DEEP CONVOLUTIONAL NEURAL NETWORK TO PREVENT ATM FRAUD BY FACIAL DISGUISE IDENTIFICATION
Suraj Nandkishor Kothawade,Sumit Baburao Tamgale
The paper proposes and demonstrates a Deep Convolutional Neural Network (DCNN) architecture to identify users with disguised face attempting a fraudulent ATM transaction. The recent introduction of Disguised Face Identification (DFI) framework proves the applicability of deep neural networks for this very problem. All the ATMs nowadays incorporate a hidden camera in them and capture the footage of their users. However, it is impossible for the police to track down the impersonators with disguised faces from the ATM footage. The proposed deep convolutional neural network is trained to identify, in real time, whether the user in the captured image is trying to cloak his identity or not. The output of the DCNN is then reported to the ATM to take appropriate steps and prevent the swindler from completing the transaction. The network is trained using a dataset of images captured in similar situations as of an ATM. The comparatively low background clutter in the images enables the network to demonstrate high accuracy in feature extraction and classification for all the different disguises.

****
A Boo(n) for Evaluating Architecture Performance
Ondrej Bajgar,Rudolf Kadlec,and Jan Kleindienst
We point out important problems with the common practice of using the best singl e model performance for comparing deep learning architectures, and we propose a method that corrects these flaws. Each time a model is trained, one gets a diffe rent result due to random factors in the training process, which include random parameter initialization and random data shuffling. Reporting the best single mo del performance does not appropriately address this stochasticity. We propose a normalized expected best-out-of-n performance (Boo_n) as a way to correct these problems.
****
A Goal-oriented Neural Conversation Model by Self-Play
Wei Wei,Quoc V. Le,Andrew M. Dai,Li-Jia Li
Building chatbots that can accomplish goals such as booking a flight ticket is a n unsolved problem in natural language understanding. Much progress has been mad e to build conversation models using techniques such as sequence2sequence modeli ng. One challenge in applying such techniques to building goal-oriented conversa tion models is that maximum likelihood-based models are not optimized toward acc omplishing goals. Recently, many methods have been proposed to address this issu e by optimizing a reward that contains task status or outcome. However, adding t he reward optimization on the fly usually provides little guidance for language construction and the conversation model soon becomes decoupled from the language model. In this paper, we propose a new setting in goal-oriented dialogue system to tighten the gap between these two aspects by enforcing model level informati on isolation on individual models between two agents. Language construction now becomes an important part in reward optimization since it is the only way inform ation can be exchanged. We experimented our models using self-play and results s howed that our method not only beat the baseline sequence2sequence model in rewa rds but can also generate human-readable meaningful conversations of comparable quality.
****
Achieving Strong Regularization for Deep Neural Networks
Dae Hoon Park,Chiu Man Ho,Yi Chang
L1 and L2 regularizers are critical tools in machine learning due to their abili ty to simplify solutions. However, imposing strong L1 or L2 regularization with gradient descent method easily fails, and this limits the generalization ability of the underlying neural networks. To understand this phenomenon, we investigat e how and why training fails for strong regularization. Specifically, we examine how gradients change over time for different regularization strengths and provi de an analysis why the gradients diminish so fast. We find that there exists a t olerance level of regularization strength, where the learning completely fails i f the regularization strength goes beyond it. We propose a simple but novel meth od, Delayed Strong Regularization, in order to moderate the tolerance level. Exp eriment results show that our proposed approach indeed achieves strong regulariz ation for both L1 and L2 regularizers and improves both accuracy and sparsity on public data sets. Our source code is published.
****
Recurrent Relational Networks for complex relational reasoning
Rasmus Berg Palm,Ulrich Paquet,Ole Winther
Humans possess an ability to abstractly reason about objects and their interacti ons, an ability not shared with state-of-the-art deep learning models. Relationa l networks, introduced by Santoro et al. (2017), add the capacity for relational reasoning to deep neural networks, but are limited in the complexity of the rea soning tasks they can address. We introduce recurrent relational networks which increase the suite of solvable tasks to those that require an order of magnitude more steps of relational reasoning. We use recurrent relational networks to sol ve Sudoku puzzles and achieve state-of-the-art results by solving 96.6% of the h ardest Sudoku puzzles, where relational networks fail to solve any. We also appl y our model to the BaBi textual QA dataset solving 19/20 tasks which is competit

ive with state-of-the-art sparse differentiable neural computers. The recurrent relational network is a general purpose module that can augment any neural network model with the capacity to do many-step relational reasoning.
****

Balanced and Deterministic Weight-sharing Helps Network Performance
Oscar Chang,Hod Lipson
Weight-sharing plays a significant role in the success of many deep neural networks, by increasing memory efficiency and incorporating useful inductive priors about the problem into the network. But understanding how weight-sharing can be used effectively in general is a topic that has not been studied extensively. Chen et al. (2015) proposed HashedNets, which augments a multi-layer perceptron with a hash table, as a method for neural network compression. We generalize this method into a framework (ArbNets) that allows for efficient arbitrary weight-sharing, and use it to study the role of weight-sharing in neural networks. We show that common neural networks can be expressed as ArbNets with different hash functions. We also present two novel hash functions, the Dirichlet hash and the Neighborhood hash, and use them to demonstrate experimentally that balanced and deterministic weight-sharing helps with the performance of a neural network.
****

Classification and Disease Localization in Histopathology Using Only Global Labels: A Weakly-Supervised Approach
Pierre Courtiol,Eric W. Tramel,Marc Sanselme,Gilles Wainrib
Analysis of histopathology slides is a critical step for many diagnoses, and in particular in oncology where it defines the gold standard. In the case of digital histopathological analysis, highly trained pathologists must review vast whole-slide-images of extreme digital resolution (100,000^2 pixels) across multiple zoom levels in order to locate abnormal regions of cells, or in some cases single cells, out of millions. The application of deep learning to this problem is hampered not only by small sample sizes, as typical datasets contain only a few hundred samples, but also by the generation of ground-truth localized annotations for training interpretable classification and segmentation models. We propose a method for disease available during training. Even without pixel-level annotations, we are able to demonstrate performance comparable with models trained with strong annotations on the Camelyon-16 lymph node metastases detection challenge. We accomplish this through the use of pre-trained deep convolutional networks, feature embedding, as well as learning via top instances and negative evidence, a multiple instance learning technique fromatp the field of semantic segmentation and object detection.
****

Improving generalization by regularizing in L^2 function space
Ari S Benjamin,Konrad Kording
Learning rules for neural networks necessarily include some form of regularization. Most regularization techniques are conceptualized and implemented in the space of parameters. However, it is also possible to regularize in the space of functions. Here, we propose to measure networks in an L^2 Hilbert space, and test a learning rule that regularizes the distance a network can travel through L^2-space each update. This approach is inspired by the slow movement of gradient descent through parameter space as well as by the natural gradient, which can be derived from a regularization term upon functional change. The resulting learning rule, which we call Hilbert-constrained gradient descent (HCGD), is thus closely related to the natural gradient but regularizes a different and more calculable metric over the space of functions. Experiments show that the HCGD is efficient and leads to considerably better generalization.
****

On Batch Adaptive Training for Deep Learning: Lower Loss and Larger Step Size
Runyao Chen,Kun Wu,Ping Luo
Mini-batch gradient descent and its variants are commonly used in deep learning. The principle of mini-batch gradient descent is to use noisy gradient calculated on a batch to estimate the real gradient, thus balancing the computation cost per iteration and the uncertainty of noisy gradient. However, its batch size is

a fixed hyper-parameter requiring manual setting before training the neural netw
ork. Yin et al. (2017) proposed a batch adaptive stochastic gradient descent (BA
-SGD) that can dynamically choose a proper batch size as learning proceeds. We e
xtend the BA-SGD to momentum algorithm and evaluate both the BA-SGD and the batc
h adaptive momentum (BA-Momentum) on two deep learning tasks from natural langua
ge processing to image classification. Experiments confirm that batch adaptive m
ethods can achieve a lower loss compared with mini-batch methods after scanning
the same epochs of data. Furthermore, our BA-Momentum is more robust against lar
ger step sizes, in that it can dynamically enlarge the batch size to reduce the
larger uncertainty brought by larger step sizes. We also identified an interesti
ng phenomenon, batch size boom. The code implementing batch adaptive framework i
s now open source, applicable to any gradient-based optimization problems.
****

AMPNet: Asynchronous Model-Parallel Training for Dynamic Neural Networks
Alexander L. Gaunt,Matthew A. Johnson,Alan Lawrence,Maik Riechert,Daniel Tarlow,
Ryota Tomioka,Dimitrios Vytiniotis,Sam Webster

New types of compute hardware in development and entering the market hold the pr
omise of revolutionizing deep learning in a manner as profound as GPUs. However,
existing software frameworks and training algorithms for deep learning have yet
to evolve to fully leverage the capability of the new wave of silicon. In parti
cular, models that exploit structured input via complex and instance-dependent c
ontrol flow are difficult to accelerate using existing algorithms and hardware t
hat typically rely on minibatching. We present an asynchronous model-parallel (A
MP) training algorithm that is specifically motivated by training on networks of
interconnected devices. Through an implementation on multi-core CPUs, we show t
hat AMP training converges to the same accuracy as conventional synchronous trai
ning algorithms in a similar number of epochs, but utilizes the available hardwa
re more efficiently, even for small minibatch sizes, resulting in shorter overal
l training times. Our framework opens the door for scaling up a new class of dee
p learning models that cannot be efficiently trained today.
****

Avoiding degradation in deep feed-forward networks by phasing out skip-connectio
ns
Ricardo Pio Monti,Sina Tootoonian,Robin Cao
A widely observed phenomenon in deep learning is the degradation problem: increa
sing
the depth of a network leads to a decrease in performance on both test and train
ing data. Novel architectures such as ResNets and Highway networks have addresse
d this issue by introducing various flavors of skip-connections or gating mechan
isms. However, the degradation problem persists in the context of plain feed-for
ward networks. In this work we propose a simple method to address this issue. Th
e proposed method poses the learning of weights in deep networks as a constraine
d optimization problem where the presence of skip-connections is penalized by La
grange multipliers. This allows for skip-connections to be introduced during the
early stages of training and subsequently phased out in a principled manner. We
demonstrate the benefits of such an approach with experiments on MNIST, fashion
-MNIST, CIFAR-10 and CIFAR-100 where the proposed method is shown to greatly dec
rease the degradation effect (compared to plain networks) and is often competiti
ve with ResNets.
****

Neighbor-encoder
Chin-Chia Michael Yeh,Yan Zhu,Evangelos E. Papalexakis,Abdullah Mueen,Eamonn Keo
gh
We propose a novel unsupervised representation learning framework called neighbo
r-encoder in which domain knowledge can be trivially incorporated into the learn
ing process without modifying the general encoder-decoder architecture. In contr
ast to autoencoder, which reconstructs the input data, neighbor-encoder reconstr
ucts the input data's neighbors. The proposed neighbor-encoder can be considered
as a generalization of autoencoder as the input data can be treated as the near

est neighbor of itself with zero distance. By reformulating the representation learning problem as a neighbor reconstruction problem, domain knowledge can be easily incorporated with appropriate definition of similarity or distance between objects. As such, any existing similarity search algorithms can be easily integrated into our framework. Applications of other algorithms (e.g., association rule mining) in our framework is also possible since the concept of ``neighbor" is an abstraction which can be appropriately defined differently in different contexts. We have demonstrated the effectiveness of our framework in various domains, including images, time series, music, etc., with various neighbor definitions. Experimental results show that neighbor-encoder outperforms autoencoder in most scenarios we considered.
****
Multi-task Learning on MNIST Image Datasets
Po-Chen Hsieh,Chia-Ping Chen
We apply multi-task learning to image classification tasks on MNIST-like datasets. MNIST dataset has been referred to as the {\em drosophila} of machine learning and has been the testbed of many learning theories. The NotMNIST dataset and the FashionMNIST dataset have been created with the MNIST dataset as reference. In this work, we exploit these MNIST-like datasets for multi-task learning. The datasets are pooled together for learning the parameters of joint classification networks. Then the learned parameters are used as the initial parameters to retrain disjoint classification networks. The baseline recognition model are all-convolution neural networks. Without multi-task learning, the recognition accuracies for MNIST, NotMNIST and FashionMNIST are 99.56\%, 97.22\% and 94.32\% respectively. With multi-task learning to pre-train the networks, the recognition accuracies are respectively 99.70\%, 97.46\% and 95.25\%. The results re-affirm that multi-task learning framework, even with data with different genres, does lead to significant improvement.

****
Synthesizing Robust Adversarial Examples
Anish Athalye,Logan Engstrom,Andrew Ilyas,Kevin Kwok
Neural network-based classifiers parallel or exceed human-level accuracy on many common tasks and are used in practical systems. Yet, neural networks are susceptible to adversarial examples, carefully perturbed inputs that cause networks to misbehave in arbitrarily chosen ways. When generated with standard methods, these examples do not consistently fool a classifier in the physical world due to a combination of viewpoint shifts, camera noise, and other natural transformations. Adversarial examples generated using standard techniques require complete control over direct input to the classifier, which is impossible in many real-world systems.

We introduce the first method for constructing real-world 3D objects that consistently fool a neural network across a wide distribution of angles and viewpoints. We present a general-purpose algorithm for generating adversarial examples that are robust across any chosen distribution of transformations. We demonstrate its application in two dimensions, producing adversarial images that are robust to noise, distortion, and affine transformation. Finally, we apply the algorithm to produce arbitrary physical 3D-printed adversarial objects, demonstrating that our approach works end-to-end in the real world. Our results show that adversarial examples are a practical concern for real-world systems.

****
Certifying Some Distributional Robustness with Principled Adversarial Training
Aman Sinha,Hongseok Namkoong,John Duchi
Neural networks are vulnerable to adversarial examples and researchers have proposed many heuristic attack and defense mechanisms. We address this problem through the principled lens of distributionally robust optimization, which guarantees performance under adversarial input perturbations. By considering a Lagrangian penalty formulation of perturbing the underlying data distribution in a Wassers

tein ball, we provide a training procedure that augments model parameter updates with worst-case perturbations of training data. For smooth losses, our procedure provably achieves moderate levels of robustness with little computational or statistical cost relative to empirical risk minimization. Furthermore, our statistical guarantees allow us to efficiently certify robustness for the population loss. For imperceptible perturbations, our method matches or outperforms heuristic approaches.

****

Latent Constraints: Learning to Generate Conditionally from Unconditional Generative Models

Jesse Engel,Matthew Hoffman,Adam Roberts

Deep generative neural networks have proven effective at both conditional and unconditional modeling of complex data distributions. Conditional generation enables interactive control, but creating new controls often requires expensive retraining. In this paper, we develop a method to condition generation without retraining the model. By post-hoc learning latent constraints, value functions identify regions in latent space that generate outputs with desired attributes, we can conditionally sample from these regions with gradient-based optimization or amortized actor functions. Combining attribute constraints with a universal “realism” constraint, which enforces similarity to the data distribution, we generate realistic conditional images from an unconditional variational autoencoder. Further, using gradient-based optimization, we demonstrate identity-preserving transformations that make the minimal adjustment in latent space to modify the attributes of an image. Finally, with discrete sequences of musical notes, we demonstrate zero-shot conditional generation, learning latent constraints in the absence of labeled data or a differentiable reward function.

****

On Characterizing the Capacity of Neural Networks Using Algebraic Topology

William H. Guss,Ruslan Salakhutdinov

The learnability of different neural architectures can be characterized directly by computable measures of data complexity. In this paper, we reframe the problem of architecture selection as understanding how data determines the most expressive and generalizable architectures suited to that data, beyond inductive bias. After suggesting algebraic topology as a measure for data complexity, we show that the power of a network to express the topological complexity of a dataset in its decision boundary is a strictly limiting factor in its ability to generalize. We then provide the first empirical characterization of the topological capacity of neural networks. Our empirical analysis shows that at every level of dataset complexity, neural networks exhibit topological phase transitions and stratification. This observation allowed us to connect existing theory to empirically driven conjectures on the choice of architectures for a single hidden layer neural networks.

****

Matrix capsules with EM routing

Geoffrey E Hinton,Sara Sabour,Nicholas Frosst

A capsule is a group of neurons whose outputs represent different properties of the same entity. Each layer in a capsule network contains many capsules. We describe a version of capsules in which each capsule has a logistic unit to represent the presence of an entity and a 4x4 matrix which could learn to represent the relationship between that entity and the viewer (the pose). A capsule in one layer votes for the pose matrix of many different capsules in the layer above by multiplying its own pose matrix by trainable viewpoint-invariant transformation matrices that could learn to represent part-whole relationships. Each of these votes is weighted by an assignment coefficient. These coefficients are iteratively updated for each image using the Expectation-Maximization algorithm such that the output of each capsule is routed to a capsule in the layer above that receives a cluster of similar votes. The transformation matrices are trained discriminatively by backpropagating through the unrolled iterations of EM between each pair of adjacent capsule layers. On the smallNORB benchmark, capsules reduce the num

ber of test errors by 45\% compared to the state-of-the-art. Capsules also show far more resistance to white box adversarial attacks than our baseline convoluti onal neural network.
****
A Bayesian Nonparametric Topic Model with Variational Auto-Encoders
Xuefei Ning,Yin Zheng,Zhuxi Jiang,Yu Wang,Huazhong Yang,Junzhou Huang
Topic modeling of text documents is one of the most important tasks in represent ation learning. In this work, we propose iTM-VAE, which is a Bayesian nonparamet ric (BNP) topic model with variational auto-encoders. On one hand, as a BNP topi c model, iTM-VAE potentially has infinite topics and can adapt the topic number to data automatically. On the other hand, different with the other BNP topic mod els, the inference of iTM-VAE is modeled by neural networks, which has rich repr esentation capacity and can be computed in a simple feed-forward manner. Two var iants of iTM-VAE are also proposed in this paper, where iTM-VAE-Prod models the generative process in products-of-experts fashion for better performance and iTM -VAE-G places a prior over the concentration parameter such that the model can a dapt a suitable concentration parameter to data automatically. Experimental resu lts on 20News and Reuters RCV1-V2 datasets show that the proposed models outperf orm the state-of-the-arts in terms of perplexity, topic coherence and document r etrieval tasks. Moreover, the ability of adjusting the concentration parameter t o data is also confirmed by experiments.
****
A Hierarchical Model for Device Placement
Azalia Mirhoseini,Anna Goldie,Hieu Pham,Benoit Steiner,Quoc V. Le,Jeff Dean
We introduce a hierarchical model for efficient placement of computational graph s onto hardware devices, especially in heterogeneous environments with a mixture of CPUs, GPUs, and other computational devices. Our method learns to assign gra ph operations to groups and to allocate those groups to available devices. The g rouping and device allocations are learned jointly. The proposed method is train ed with policy gradient and requires no human intervention. Experiments with wid ely-used
computer vision and natural language models show that our algorithm can find opt imized, non-trivial placements for TensorFlow computational graphs with over 80, 000 operations. In addition, our approach outperforms placements by human
experts as well as a previous state-of-the-art placement method based on deep re inforcement learning. Our method achieves runtime reductions of up to 60.6% per training step when applied to models such as Neural Machine Translation.
****
Apprentice: Using Knowledge Distillation Techniques To Improve Low-Precision Net work Accuracy
Asit Mishra,Debbie Marr
Deep learning networks have achieved state-of-the-art accuracies on computer vis ion workloads like image classification and object detection. The performant sys tems, however, typically involve big models with numerous parameters. Once train ed, a challenging aspect for such top performing models is deployment on resourc e constrained inference systems -- the models (often deep networks or wide netwo rks or both) are compute and memory intensive. Low precision numerics and model compression using knowledge distillation are popular techniques to lower both th e compute requirements and memory footprint of these deployed models. In this pa per, we study the combination of these two techniques and show that the performa nce of low precision networks can be significantly improved by using knowledge d istillation techniques. We call our approach Apprentice and show state-of-the-ar t accuracies using ternary precision and 4-bit precision for many variants of Re sNet architecture on ImageNet dataset. We study three schemes in which one can a pply knowledge distillation techniques to various stages of the train-and-deploy pipeline.
****
Residual Connections Encourage Iterative Inference
Stanis■aw Jastrzebski,Devansh Arpit,Nicolas Ballas,Vikas Verma,Tong Che,Yoshua B engio

Residual networks (Resnets) have become a prominent architecture in deep learning. However, a comprehensive understanding of Resnets is still a topic of ongoing research. A recent view argues that Resnets perform iterative refinement of features. We attempt to further expose properties of this aspect. To this end, we study Resnets both analytically and empirically. We formalize the notion of iterative refinement in Resnets by showing that residual architectures naturally encourage features to move along the negative gradient of loss during the feedforward phase. In addition, our empirical analysis suggests that Resnets are able to perform both representation learning and iterative refinement. In general, a Resnet block tends to concentrate representation learning behavior in the first few layers while higher layers perform iterative refinement of features. Finally we observe that sharing residual layers naively leads to representation explosion and hurts generalization performance, and show that simple existing strategies can help alleviating this problem.
****
Gradients explode - Deep Networks are shallow - ResNet explained
George Philipp,Dawn Song,Jaime G. Carbonell
Whereas it is believed that techniques such as Adam, batch normalization and, more recently, SeLU nonlinearities ``solve'' the exploding gradient problem, we show that this is not the case and that in a range of popular MLP architectures, exploding gradients exist and that they limit the depth to which networks can be effectively trained, both in theory and in practice. We explain why exploding gradients occur and highlight the {\it collapsing domain problem}, which can arise in architectures that avoid exploding gradients.

ResNets have significantly lower gradients and thus can circumvent the exploding gradient problem, enabling the effective training of much deeper networks, which we show is a consequence of a surprising mathematical property. By noticing that {\it any neural network is a residual network}, we devise the {\it residual trick}, which reveals that introducing skip connections simplifies the network mathematically, and that this simplicity may be the major cause for their success.
****
Network of Graph Convolutional Networks Trained on Random Walks
Sami Abu-El-Haija,Amol Kapoor,Bryan Perozzi,Joonseok Lee
Graph Convolutional Networks (GCNs) are a recently proposed architecture which has had success in semi-supervised learning on graph-structured data. At the same time, unsupervised learning of graph embeddings has benefited from the information contained in random walks. In this paper we propose a model, Network of GCNs (N-GCN), which marries these two lines of work. At its core, N-GCN trains multiple instances of GCNs over node pairs discovered at different distances in random walks, and learns a combination of the instance outputs which optimizes the classification objective. Our experiments show that our proposed N-GCN model achieves state-of-the-art performance on all of the challenging node classification tasks we consider: Cora, Citeseer, Pubmed, and PPI. In addition, our proposed method has other desirable properties, including generalization to recently proposed semi-supervised learning methods such as GraphSAGE, allowing us to propose N-SAGE, and resilience to adversarial input perturbations.
****
Training Autoencoders by Alternating Minimization
Sneha Kudugunta,Adepu Shankar,Surya Chavali,Vineeth Balasubramanian,Purushottam Kar
We present DANTE, a novel method for training neural networks, in particular autoencoders, using the alternating minimization principle. DANTE provides a distinct perspective in lieu of traditional gradient-based backpropagation techniques commonly used to train deep networks. It utilizes an adaptation of quasi-convex optimization techniques to cast autoencoder training as a bi-quasi-convex optimization problem. We show that for autoencoder configurations with both differentiable (e.g. sigmoid) and non-differentiable (e.g. ReLU) activation functions, we can perform the alternations very effectively. DANTE effortlessly extends to networks with multiple hidden layers and varying network configurations. In experim

ents on standard datasets, autoencoders trained using the proposed method were found to be very promising when compared to those trained using traditional backpropagation techniques, both in terms of training speed, as well as feature extraction and reconstruction performance.
****

STRUCTURED ALIGNMENT NETWORKS

Yang Liu,Matt Gardner

Many tasks in natural language processing involve comparing two sentences to compute some notion of relevance, entailment, or similarity. Typically this comparison is done either at the word level or at the sentence level, with no attempt to leverage the inherent structure of the sentence. When sentence structure is used for comparison, it is obtained during a non-differentiable pre-processing step, leading to propagation of errors. We introduce a model of structured alignments between sentences, showing how to compare two sentences by matching their latent structures. Using a structured attention mechanism, our model matches possible spans in the first sentence to possible spans in the second sentence, simultaneously discovering the tree structure of each sentence and performing a comparison, in a model that is fully differentiable and is trained only on the comparison objective. We evaluate this model on two sentence comparison tasks: the Stanford natural language inference dataset and the TREC-QA dataset. We find that comparing spans results in superior performance to comparing words individually, and that the learned trees are consistent with actual linguistic structures.
****

Improve Training Stability of Semi-supervised Generative Adversarial Networks with Collaborative Training

Dalei Wu,Xiaohua Liu

Improved generative adversarial network (Improved GAN) is a successful method of using generative adversarial models to solve the problem of semi-supervised learning. However, it suffers from the problem of unstable training. In this paper, we found that the instability is mostly due to the vanishing gradients on the generator. To remedy this issue, we propose a new method to use collaborative training to improve the stability of semi-supervised GAN with the combination of Wasserstein GAN. The experiments have shown that our proposed method is more stable than the original Improved GAN and achieves comparable classification accuracy on different data sets.
****

Twin Networks: Matching the Future for Sequence Generation

Dmitriy Serdyuk,Nan Rosemary Ke,Alessandro Sordoni,Adam Trischler,Chris Pal,Yoshua Bengio

We propose a simple technique for encouraging generative RNNs to plan ahead. We train a ``backward'' recurrent network to generate a given sequence in reverse order, and we encourage states of the forward model to predict cotemporal states of the backward model. The backward network is used only during training, and plays no role during sampling or inference. We hypothesize that our approach eases modeling of long-term dependencies by implicitly forcing the forward states to hold information about the longer-term future (as contained in the backward states). We show empirically that our approach achieves 9% relative improvement for a speech recognition task, and achieves significant improvement on a COCO caption generation task.
****

Linearly Constrained Weights: Resolving the Vanishing Gradient Problem by Reducing Angle Bias

Takuro Kutsuna

In this paper, we first identify \textit{angle bias}, a simple but remarkable phenomenon that causes the vanishing gradient problem in a multilayer perceptron (MLP) with sigmoid activation functions. We then propose \textit{linearly constrained weights (LCW)} to reduce the angle bias in a neural network, so as to train the network under the constraints that the sum of the elements of each weight vector is zero. A reparameterization technique is presented to efficiently train a model with LCW by embedding the constraints on weight vectors into the structu

re of the network. Interestingly, batch normalization (Ioffe & Szegedy, 2015) can be viewed as a mechanism to correct angle bias. Preliminary experiments show that LCW helps train a 100-layered MLP more efficiently than does batch normalization.
****

Topic-Based Question Generation

Wenpeng Hu,Bing Liu,Rui Yan,Dongyan Zhao,Jinwen Ma

Asking questions is an important ability for a chatbot. This paper focuses on question generation. Although there are existing works on question generation based on a piece of descriptive text, it remains to be a very challenging problem. In the paper, we propose a new question generation problem, which also requires the input of a target topic in addition to a piece of descriptive text. The key reason for proposing the new problem is that in practical applications, we found that useful questions need to be targeted toward some relevant topics. One almost never asks a random question in a conversation. Due to the fact that given a descriptive text, it is often possible to ask many types of questions, generating a question without knowing what it is about is of limited use. To solve the problem, we propose a novel neural network that is able to generate topic-specific questions. One major advantage of this model is that it can be trained directly using a question-answering corpus without requiring any additional annotations like annotating topics in the questions or answers. Experimental results show that our model outperforms the state-of-the-art baseline.
****

Stabilizing GAN Training with Multiple Random Projections

Behnam Neyshabur,Srinadh Bhojanapalli,Ayan Chakrabarti

Training generative adversarial networks is unstable in high-dimensions as the true data distribution tends to be concentrated in a small fraction of the ambient space. The discriminator is then quickly able to classify nearly all generated samples as fake, leaving the generator without meaningful gradients and causing it to deteriorate after a point in training. In this work, we propose training a single generator simultaneously against an array of discriminators, each of which looks at a different random low-dimensional projection of the data. Individual discriminators, now provided with restricted views of the input, are unable to reject generated samples perfectly and continue to provide meaningful gradients to the generator throughout training. Meanwhile, the generator learns to produce samples consistent with the full data distribution to satisfy all discriminators simultaneously. We demonstrate the practical utility of this approach experimentally, and show that it is able to produce image samples with higher quality than traditional training with a single discriminator.
****

Unsupervised Deep Structure Learning by Recursive Dependency Analysis

Raanan Y. Yehezkel Rohekar,Guy Koren,Shami Nisimov,Gal Novik

We introduce an unsupervised structure learning algorithm for deep, feed-forward, neural networks. We propose a new interpretation for depth and inter-layer connectivity where a hierarchy of independencies in the input distribution is encoded in the network structure. This results in structures allowing neurons to connect to neurons in any deeper layer skipping intermediate layers. Moreover, neurons in deeper layers encode low-order (small condition sets) independencies and have a wide scope of the input, whereas neurons in the first layers encode higher-order (larger condition sets) independencies and have a narrower scope. Thus, the depth of the network is automatically determined---equal to the maximal order of independence in the input distribution, which is the recursion-depth of the algorithm. The proposed algorithm constructs two main graphical models: 1) a generative latent graph (a deep belief network) learned from data and 2) a deep discriminative graph constructed from the generative latent graph. We prove that conditional dependencies between the nodes in the learned generative latent graph are preserved in the class-conditional discriminative graph. Finally, a deep neural network structure is constructed based on the discriminative graph. We demonstrate on image classification benchmarks that the algorithm replaces the deepest layers (convolutional and dense layers) of common convolutional networks, achi

eving high classification accuracy, while constructing significantly smaller structures. The proposed structure learning algorithm requires a small computational cost and runs efficiently on a standard desktop CPU.
****
Towards Image Understanding from Deep Compression Without Decoding
Robert Torfason,Fabian Mentzer,Eirikur Agustsson,Michael Tschannen,Radu Timofte,Luc Van Gool
Motivated by recent work on deep neural network (DNN)-based image compression methods showing potential improvements in image quality, savings in storage, and bandwidth reduction, we propose to perform image understanding tasks such as classification and segmentation directly on the compressed representations produced by these compression methods. Since the encoders and decoders in DNN-based compression methods are neural networks with feature-maps as internal representations of the images, we directly integrate these with architectures for image understanding. This bypasses decoding of the compressed representation into RGB space and reduces computational cost. Our study shows that accuracies comparable to networks that operate on compressed RGB images can be achieved while reducing the computational complexity up to $2\times$. Furthermore, we show that synergies are obtained by jointly training compression networks with classification networks on the compressed representations, improving image quality, classification accuracy, and segmentation performance. We find that inference from compressed representations is particularly advantageous compared to inference from compressed RGB images for aggressive compression rates.
****
Mixed Precision Training
Paulius Micikevicius,Sharan Narang,Jonah Alben,Gregory Diamos,Erich Elsen,David Garcia,Boris Ginsburg,Michael Houston,Oleksii Kuchaiev,Ganesh Venkatesh,Hao Wu
Increasing the size of a neural network typically improves accuracy but also increases the memory and compute requirements for training the model. We introduce methodology for training deep neural networks using half-precision floating point numbers, without losing model accuracy or having to modify hyper-parameters. This nearly halves memory requirements and, on recent GPUs, speeds up arithmetic. Weights, activations, and gradients are stored in IEEE half-precision format. Since this format has a narrower range than single-precision we propose three techniques for preventing the loss of critical information. Firstly, we recommend maintaining a single-precision copy of weights that accumulates the gradients after each optimizer step (this copy is rounded to half-precision for the forward- and back-propagation). Secondly, we propose loss-scaling to preserve gradient values with small magnitudes. Thirdly, we use half-precision arithmetic that accumulates into single-precision outputs, which are converted to half-precision before storing to memory. We demonstrate that the proposed methodology works across a wide variety of tasks and modern large scale (exceeding 100 million parameters) model architectures, trained on large datasets.
****
Revisiting Bayes by Backprop
Meire Fortunato,Charles Blundell,Oriol Vinyals
In this work we explore a straightforward variational Bayes scheme for Recurrent Neural Networks.
Firstly, we show that a simple adaptation of truncated backpropagation through time can yield good quality uncertainty estimates and superior regularisation at only a small extra computational cost during training, also reducing the amount of parameters by 80\%.
Secondly, we demonstrate how a novel kind of posterior approximation yields further improvements to the performance of Bayesian RNNs. We incorporate local gradient information into the approximate posterior to sharpen it around the current batch statistics. We show how this technique is not exclusive to recurrent neural networks and can be applied more widely to train Bayesian neural networks.
We also empirically demonstrate how Bayesian RNNs are superior to traditional RNNs on a language modelling benchmark and an image captioning task, as well as showing how each of these methods improve our model over a variety of other scheme

s for training them. We also introduce a new benchmark for studying uncertainty for language models so future methods can be easily compared.
****

Relational Neural Expectation Maximization: Unsupervised Discovery of Objects and their Interactions
Sjoerd van Steenkiste,Michael Chang,Klaus Greff,Jürgen Schmidhuber
Common-sense physical reasoning is an essential ingredient for any intelligent agent operating in the real-world. For example, it can be used to simulate the environment, or to infer the state of parts of the world that are currently unobserved. In order to match real-world conditions this causal knowledge must be learned without access to supervised data. To address this problem we present a novel method that learns to discover objects and model their physical interactions from raw visual images in a purely unsupervised fashion. It incorporates prior knowledge about the compositional nature of human perception to factor interactions between object-pairs and learn efficiently. On videos of bouncing balls we show the superior modelling capabilities of our method compared to other unsupervised neural approaches that do not incorporate such prior knowledge. We demonstrate its ability to handle occlusion and show that it can extrapolate learned knowledge to scenes with different numbers of objects.
****

Long-term Forecasting using Tensor-Train RNNs
Rose Yu,Stephan Zheng,Anima Anandkumar,Yisong Yue
We present Tensor-Train RNN (TT-RNN), a novel family of neural sequence architectures for multivariate forecasting in environments with nonlinear dynamics. Long-term forecasting in such systems is highly challenging, since there exist long-term temporal dependencies, higher-order correlations and sensitivity to error propagation. Our proposed tensor recurrent architecture addresses these issues by learning the nonlinear dynamics directly using higher order moments and high-order state transition functions. Furthermore, we decompose the higher-order structure using the tensor-train (TT) decomposition to reduce the number of parameters while preserving the model performance. We theoretically establish the approximation properties of Tensor-Train RNNs for general sequence inputs, and such guarantees are not available for usual RNNs. We also demonstrate significant long-term prediction improvements over general RNN and LSTM architectures on a range of simulated environments with nonlinear dynamics, as well on real-world climate and traffic data.
****

BLOCK-DIAGONAL HESSIAN-FREE OPTIMIZATION FOR TRAINING NEURAL NETWORKS
Huishuai Zhang,Caiming Xiong,James Bradbury,Richard Socher
Second-order methods for neural network optimization have several advantages over methods based on first-order gradient descent, including better scaling to large mini-batch sizes and fewer updates needed for convergence. But they are rarely applied to deep learning in practice because of high computational cost and the need for model-dependent algorithmic variations. We introduce a vari- ant of the Hessian-free method that leverages a block-diagonal approximation of the generalized Gauss-Newton matrix. Our method computes the curvature approximation matrix only for pairs of parameters from the same layer or block of the neural network and performs conjugate gradient updates independently for each block. Experiments on deep autoencoders, deep convolutional networks, and multilayer LSTMs demonstrate better convergence and generalization compared to the original Hessian-free approach and the Adam method.
****

Learning Deep Mean Field Games for Modeling Large Population Behavior
Jiachen Yang,Xiaojing Ye,Rakshit Trivedi,Huan Xu,Hongyuan Zha
We consider the problem of representing collective behavior of large populations and predicting the evolution of a population distribution over a discrete state space. A discrete time mean field game (MFG) is motivated as an interpretable model founded on game theory for understanding the aggregate effect of individual actions and predicting the temporal evolution of population distributions. We achieve a synthesis of MFG and Markov decision processes (MDP) by showing that a

special MFG is reducible to an MDP. This enables us to broaden the scope of mean field game theory and infer MFG models of large real-world systems via deep inverse reinforcement learning. Our method learns both the reward function and forward dynamics of an MFG from real data, and we report the first empirical test of a mean field game model of a real-world social media population.
****

Continuous-fidelity Bayesian Optimization with Knowledge Gradient
Jian Wu,Peter I. Frazier
While Bayesian optimization (BO) has achieved great success in optimizing expensive-to-evaluate black-box functions, especially tuning hyperparameters of neural networks, methods such as random search (Li et al., 2016) and multi-fidelity BO (e.g. Klein et al. (2017)) that exploit cheap approximations, e.g. training on a smaller training data or with fewer iterations, can outperform standard BO approaches that use only full-fidelity observations. In this paper, we propose a novel Bayesian optimization algorithm, the continuous-fidelity knowledge gradient (cfKG) method, that can be used when fidelity is controlled by one or more continuous settings such as training data size and the number of training iterations. cfKG characterizes the value of the information gained by sampling a point at a given fidelity, choosing to sample at the point and fidelity with the largest value per unit cost. Furthermore, cfKG can be generalized, following Wu et al. (2017), to settings where derivatives are available in the optimization process, e.g. large-scale kernel learning, and where more than one point can be evaluated simultaneously. Numerical experiments show that cfKG outperforms state-of-art algorithms when optimizing synthetic functions, tuning convolutional neural networks (CNNs) on CIFAR-10 and SVHN, and in large-scale kernel learning.
****

Semi-Supervised Learning via New Deep Network Inversion
Balestriero R.,Roger V.,Glotin H.,Baraniuk R.
We exploit a recently derived inversion scheme for arbitrary deep neural networks to develop a new semi-supervised learning framework that applies to a wide range of systems and problems.
The approach reaches current state-of-the-art methods on MNIST and provides reasonable performances on SVHN and CIFAR10. Through the introduced method, residual networks are for the first time applied to semi-supervised tasks. Experiments with one-dimensional signals highlight the generality of the method. Importantly, our approach is simple, efficient, and requires no change in the deep network architecture.
****

Phase Conductor on Multi-layered Attentions for Machine Comprehension
Rui Liu,Wei Wei,Weiguang Mao,Maria Chikina
Attention models have been intensively studied to improve NLP tasks such as machine comprehension via both question-aware passage attention model and self-matching attention model. Our research proposes phase conductor (PhaseCond) for attention models in two meaningful ways. First, PhaseCond, an architecture of multi-layered attention models, consists of multiple phases each implementing a stack of attention layers producing passage representations and a stack of inner or outer fusion layers regulating the information flow. Second, we extend and improve the dot-product attention function for PhaseCond by simultaneously encoding multiple question and passage embedding layers from different perspectives. We demonstrate the effectiveness of our proposed model PhaseCond on the SQuAD dataset, showing that our model significantly outperforms both state-of-the-art single-layered and multiple-layered attention models. We deepen our results with new findings via both detailed qualitative analysis and visualized examples showing the dynamic changes through multi-layered attention models.
****

Data-driven Feature Sampling for Deep Hyperspectral Classification and Segmentation
William M. Severa,Jerilyn A. Timlin,Suraj Kholwadwala,Conrad D. James,James B. Aimone
The high dimensionality of hyperspectral imaging forces unique challenges in sco

pe, size and processing requirements. Motivated by the potential for an in-the-
field cell sorting detector, we examine a Synechocystis sp. PCC 6803 dataset whe
rein cells are grown alternatively in nitrogen rich or deplete cultures. We use
 deep learning techniques to both successfully classify cells and generate a mas
k segmenting the cells/condition from the background. Further, we use the classi
fication accuracy to guide a data-driven, iterative feature selection method, al
lowing the design neural networks requiring 90% fewer input features with little
 accuracy degradation.
****
Meta-Learning Transferable Active Learning Policies by Deep Reinforcement Learni
ng
Kunkun Pang,Mingzhi Dong,Timothy Hospedales
Active learning (AL) aims to enable training high performance classifiers with l
ow annotation cost by predicting which subset of unlabelled instances would be m
ost beneficial to label. The importance of AL has motivated extensive research,
proposing a wide variety of manually designed AL algorithms with diverse theoret
ical and intuitive motivations. In contrast to this body of research, we propose
 to treat active learning algorithm design as a meta-learning problem and learn
the best criterion from data. We model an active learning algorithm as a deep ne
ural network that inputs the base learner state and the unlabelled point set and
 predicts the best point to annotate next. Training this active query policy net
work with reinforcement learning, produces the best non-myopic policy for a give
n dataset. The key challenge in achieving a general solution to AL then becomes
that of learner generalisation, particularly across heterogeneous datasets. We p
ropose a multi-task dataset-embedding approach that allows dataset-agnostic acti
ve learners to be trained. Our evaluation shows that AL algorithms trained in th
is way can directly generalize across diverse problems.
****
Learning to Compute Word Embeddings On the Fly
Dzmitry Bahdanau,Tom Bosc,Stanis■aw Jastrz■bski,Edward Grefenstette,Pascal Vince
nt,Yoshua Bengio
Words in natural language follow a Zipfian distribution whereby some words are f
requent but most are rare. Learning representations for words in the ``long tail
'' of this distribution requires enormous amounts of data.
Representations of rare words trained directly on end tasks are usually poor, re
quiring us to pre-train embeddings on external data, or treat all rare words as
out-of-vocabulary words with a unique representation. We provide a method for pr
edicting embeddings of rare words on the fly from small amounts of auxiliary dat
a with a network trained end-to-end for the downstream task. We show that this i
mproves results against baselines where embeddings are trained on the end task f
or reading comprehension, recognizing textual entailment and language modeling.

****
Preliminary theoretical troubleshooting in Variational Autoencoder
Shiqi Liu,Qian Zhao,Xiangyong Cao,Deyu Meng,Zilu Ma,Tao Yu
What would be learned by variational autoencoder(VAE) and what influence the dis
entanglement of VAE? This paper tries to preliminarily address VAE's intrinsic d
imension, real factor, disentanglement and indicator issues theoretically in the
 idealistic situation and implementation issue practically through noise modelin
g perspective in the realistic case. On intrinsic dimension issue, due to infor
mation conservation, the idealistic VAE learns and only learns intrinsic factor
dimension. Besides, suggested by mutual information separation property, the con
straint induced by Gaussian prior to the VAE objective encourages the informatio
n sparsity in dimension. On disentanglement issue,   subsequently, inspired by i
nformation conservation theorem the clarification on disentanglement in this pap
er is made. On real factor issue, due to factor equivalence, the idealistic VAE
possibly learns any factor set in the equivalence class. On indicator issue, th
e behavior of current disentanglement metric is discussed, and several performan
ce indicators regarding the disentanglement and generating influence are subsequ
ently raised to evaluate the performance of VAE model and to supervise the used

factors. On implementation issue, the experiments under noise modeling and constraints empirically testify the theoretical analysis and also show their own characteristic in pursuing disentanglement.
****

Learning Generative Models with Locally Disentangled Latent Factors
Brady Neal,Alex Lamb,Sherjil Ozair,Devon Hjelm,Aaron Courville,Yoshua Bengio,Ioannis Mitliagkas
One of the most successful techniques in generative models has been decomposing a complicated generation task into a series of simpler generation tasks. For example, generating an image at a low resolution and then learning to refine that into a high resolution image often improves results substantially. Here we explore a novel strategy for decomposing generation for complicated objects in which we first generate latent variables which describe a subset of the observed variables, and then map from these latent variables to the observed space. We show that this allows us to achieve decoupled training of complicated generative models and present both theoretical and experimental results supporting the benefit of such an approach.
****

Hybed: Hyperbolic Neural Graph Embedding
Benjamin Paul Chamberlain,James R Clough,Marc Peter Deisenroth
Neural embeddings have been used with great success in Natural Language Processing (NLP) where they provide compact representations that encapsulate word similarity and attain state-of-the-art performance in a range of linguistic tasks. The success of neural embeddings has prompted significant amounts of research into applications in domains other than language. One such domain is graph-structured data, where embeddings of vertices can be learned that encapsulate vertex similarity and improve performance on tasks including edge prediction and vertex labelling. For both NLP and graph-based tasks, embeddings in high-dimensional Euclidean spaces have been learned.
However, recent work has shown that the appropriate isometric space for embedding complex networks is not the flat Euclidean space, but a negatively curved hyperbolic space. We present a new concept that exploits these recent insights and propose learning neural embeddings of graphs in hyperbolic space. We provide experimental evidence that hyperbolic embeddings significantly outperform Euclidean embeddings on vertex classification tasks for several real-world public datasets.
****

Model-based imitation learning from state trajectories
Subhajit Chaudhury,Daiki Kimura,Tadanobu Inoue,Ryuki Tachibana
Imitation learning from demonstrations usually relies on learning a policy from trajectories of optimal states and actions. However, in real life expert demonstrations, often the action information is missing and only state trajectories are available. We present a model-based imitation learning method that can learn environment-specific optimal actions only from expert state trajectories. Our proposed method starts with a model-free reinforcement learning algorithm with a heuristic reward signal to sample environment dynamics, which is then used to train the state-transition probability. Subsequently, we learn the optimal actions from expert state trajectories by supervised learning, while back-propagating the error gradients through the modeled environment dynamics. Experimental evaluations show that our proposed method successfully achieves performance similar to (state, action) trajectory-based traditional imitation learning methods even in the absence of action information, with much fewer iterations compared to conventional model-free reinforcement learning methods. We also demonstrate that our method can learn to act from only video demonstrations of expert agent for simple games and can learn to achieve desired performance in less number of iterations.
****

Learning Intrinsic Sparse Structures within Long Short-Term Memory
Wei Wen,Yuxiong He,Samyam Rajbhandari,Minjia Zhang,Wenhan Wang,Fang Liu,Bin Hu,Yiran Chen,Hai Li
Model compression is significant for the wide adoption of Recurrent Neural Netwo

rks (RNNs) in both user devices possessing limited resources and business clusters requiring quick responses to large-scale service requests. This work aims to learn structurally-sparse Long Short-Term Memory (LSTM) by reducing the sizes of basic structures within LSTM units, including input updates, gates, hidden states, cell states and outputs. Independently reducing the sizes of basic structures can result in inconsistent dimensions among them, and consequently, end up with invalid LSTM units. To overcome the problem, we propose Intrinsic Sparse Structures (ISS) in LSTMs. Removing a component of ISS will simultaneously decrease the sizes of all basic structures by one and thereby always maintain the dimension consistency. By learning ISS within LSTM units, the obtained LSTMs remain regular while having much smaller basic structures. Based on group Lasso regularization, our method achieves 10.59x speedup without losing any perplexity of a language modeling of Penn TreeBank dataset. It is also successfully evaluated through a compact model with only 2.69M weights for machine Question Answering of SQuAD dataset. Our approach is successfully extended to non- LSTM RNNs, like Recurrent Highway Networks (RHNs). Our source code is available.
****
Efficient Exploration through Bayesian   Deep Q-Networks
Kamyar Azizzadenesheli,Emma Brunskill,Animashree Anandkumar
We propose Bayesian Deep Q-Network (BDQN), a practical Thompson sampling based Reinforcement Learning (RL) Algorithm. Thompson sampling allows for targeted exploration in high dimensions through posterior sampling but is usually computationally expensive. We address this limitation by introducing uncertainty only at the output layer of the network through a Bayesian Linear Regression (BLR) model, which can be trained with fast closed-form updates and its samples can be drawn efficiently through the Gaussian distribution. We apply our method to a wide range of Atari Arcade Learning Environments. Since BDQN carries out more efficient exploration, it is able to reach higher rewards substantially faster than a key baseline, DDQN.
****
Empirical Risk Landscape Analysis for Understanding Deep Neural Networks
Pan Zhou,Jiashi Feng
This work aims to provide comprehensive landscape analysis of empirical risk in deep neural networks (DNNs), including the convergence behavior of its gradient, its stationary points and the empirical risk itself to their corresponding population counterparts, which reveals how various network parameters determine the convergence performance. In particular, for an l-layer linear neural network consisting of $\dm_i$ neurons in the i-th layer, we prove the gradient of its empirical risk uniformly converges to the one of its population risk, at the rate of $\mathcal{O}(r^{2l} \sqrt{l\sqrt{\max_i \dm_i} s\log(d/l)/n})$. Here d is the total weight dimension, s is the number of nonzero entries of all the weights and the magnitude of weights per layer is upper bounded by r. Moreover, we prove the one-to-one correspondence of the non-degenerate stationary points between the empirical and population risks and provide convergence guarantee for each pair. We also establish the uniform convergence of the empirical risk to its population counterpart and further derive the stability and generalization bounds for the empirical risk. In addition, we analyze these properties for deep \emph{nonlinear} neural networks with sigmoid activation functions. We prove similar results for convergence behavior of their empirical risk gradients, non-degenerate stationary points as well as the empirical risk itself.

To our best knowledge, this work is the first one theoretically characterizing the uniform convergence of the gradient and stationary points of the empirical risk of DNN models, which benefits the theoretical understanding on how the neural network depth l, the layer width $\dm_i$, the network size d, the sparsity in weight and the parameter magnitude r determine the neural network landscape.
****
Training Neural Machines with Partial Traces
Matthew Mirman,Dimitar Dimitrov,Pavle Djordjevich,Timon Gehr,Martin Vechev

We present a novel approach for training neural abstract architectures which incorporates (partial) supervision over the machine's interpretable components. To cleanly capture the set of neural architectures to which our method applies, we introduce the concept of a differential neural computational machine (∂NCM) and show that several existing architectures (e.g., NTMs, NRAMs) can be instantiated as a ∂NCM and can thus benefit from any amount of additional supervision over their interpretable components. Based on our method, we performed a detailed experimental evaluation with both, the NTM and NRAM architectures, and showed that the approach leads to significantly better convergence and generalization capabilities of the learning phase than when training using only input-output examples.

****

A DIRT-T Approach to Unsupervised Domain Adaptation

Rui Shu,Hung Bui,Hirokazu Narui,Stefano Ermon

Domain adaptation refers to the problem of leveraging labeled data in a source domain to learn an accurate model in a target domain where labels are scarce or unavailable. A recent approach for finding a common representation of the two domains is via domain adversarial training (Ganin & Lempitsky, 2015), which attempts to induce a feature extractor that matches the source and target feature distributions in some feature space. However, domain adversarial training faces two critical limitations: 1) if the feature extraction function has high-capacity, then feature distribution matching is a weak constraint, 2) in non-conservative domain adaptation (where no single classifier can perform well in both the source and target domains), training the model to do well on the source domain hurts performance on the target domain. In this paper, we address these issues through the lens of the cluster assumption, i.e., decision boundaries should not cross high-density data regions. We propose two novel and related models: 1) the Virtual Adversarial Domain Adaptation (VADA) model, which combines domain adversarial training with a penalty term that punishes the violation the cluster assumption; 2) the Decision-boundary Iterative Refinement Training with a Teacher (DIRT-T) model, which takes the VADA model as initialization and employs natural gradient steps to further minimize the cluster assumption violation. Extensive empirical results demonstrate that the combination of these two models significantly improve the state-of-the-art performance on the digit, traffic sign, and Wi-Fi recognition domain adaptation benchmarks.

****

Revisiting The Master-Slave Architecture In Multi-Agent Deep Reinforcement Learning

Xiangyu Kong,Fangchen Liu,Bo Xin,Yizhou Wang

Many tasks in artificial intelligence require the collaboration of multiple agents. We exam deep reinforcement learning for multi-agent domains. Recent research efforts often take the form of two seemingly conflicting perspectives, the decentralized perspective, where each agent is supposed to have its own controller; and the centralized perspective, where one assumes there is a larger model controlling all agents. In this regard, we revisit the idea of the master-slave architecture by incorporating both perspectives within one framework. Such a hierarchical structure naturally leverages advantages from one another. The idea of combining both perspective is intuitive and can be well motivated from many real world systems, however, out of a variety of possible realizations, we highlights three key ingredients, i.e. composed action representation, learnable communication and independent reasoning. With network designs to facilitate these explicitly, our proposal consistently outperforms latest competing methods both in synthetics experiments and when applied to challenging StarCraft micromanagement tasks.

****

Zero-Shot Visual Imitation

Deepak Pathak,Parsa Mahmoudieh,Guanghao Luo,Pulkit Agrawal,Dian Chen,Yide Shentu,Evan Shelhamer,Jitendra Malik,Alexei A. Efros,Trevor Darrell

The current dominant paradigm for imitation learning relies on strong supervisio

n of expert actions to learn both 'what' and 'how' to imitate. We pursue an alte rnative paradigm wherein an agent first explores the world without any expert su pervision and then distills its experience into a goal-conditioned skill policy with a novel forward consistency loss. In our framework, the role of the expert is only to communicate the goals (i.e., what to imitate) during inference. The l earned policy is then employed to mimic the expert (i.e., how to imitate) after seeing just a sequence of images demonstrating the desired task. Our method is ' zero-shot' in the sense that the agent never has access to expert actions during training or for the task demonstration at inference. We evaluate our zero-shot imitator in two real-world settings: complex rope manipulation with a Baxter rob ot and navigation in previously unseen office environments with a TurtleBot. Thr ough further experiments in VizDoom simulation, we provide evidence that better mechanisms for exploration lead to learning a more capable policy which in turn improves end task performance. Videos, models, and more details are available at https://pathak22.github.io/zeroshot-imitation/.
****

Bayesian Hypernetworks
David Krueger,Chin-Wei Huang,Riashat Islam,Ryan Turner,Alexandre Lacoste,Aaron C ourville
We propose Bayesian hypernetworks: a framework for approximate Bayesian inferenc e in neural networks. A Bayesian hypernetwork, h, is a neural network which lear ns to transform a simple noise distribution, p(e) = N(0,I), to a distribution q(t) := q(h(e)) over the parameters t of another neural network (the ``primary net work). We train q with variational inference, using an invertible h to enable ef ficient estimation of the variational lower bound on the posterior p(t | D) via sampling. In contrast to most methods for Bayesian deep learning, Bayesian hyper nets can represent a complex multimodal approximate posterior with correlations between parameters, while enabling cheap iid sampling of q(t). In practice, Bay esian hypernets provide a better defense against adversarial examples than dropo ut, and also exhibit competitive performance on a suite of tasks which evaluate model uncertainty, including regularization, active learning, and anomaly detect ion.

****

Using Deep Reinforcement Learning to Generate Rationales for Molecules
Benson Chen,Connor Coley,Regina Barzilay,Tommi Jaakkola
Deep learning algorithms are increasingly used in modeling chemical processes. H owever, black box predictions without rationales have limited used in practical applications, such as drug design. To this end, we learn to identify molecular s ubstructures -- rationales -- that are associated with the target chemical prope rty (e.g., toxicity). The rationales are learned in an unsupervised fashion, req uiring no additional information beyond the end-to-end task. We formulate this p roblem as a reinforcement learning problem over the molecular graph, parametrize d by two convolution networks corresponding to the rationale selection and predi ction based on it, where the latter induces the reward function. We evaluate the approach on two benchmark toxicity datasets. We demonstrate that our model sust ains high performance under the additional constraint that predictions strictly follow the rationales. Additionally, we validate the extracted rationales throug h comparison against those described in chemical literature and through syntheti c experiments.
****

Directing Generative Networks with Weighted Maximum Mean Discrepancy
Maurice Diesendruck,Guy W. Cole,Sinead Williamson
The maximum mean discrepancy (MMD) between two probability measures P
and Q is a metric that is zero if and only if all moments of the two measures
are equal, making it an appealing statistic for two-sample tests. Given i.i.d. s amples
from P and Q, Gretton et al. (2012) show that we can construct an unbiased
estimator for the square of the MMD between the two distributions. If P is a
distribution of interest and Q is the distribution implied by a generative neura

l
network with stochastic inputs, we can use this estimator to train our neural ne twork.
However, in practice we do not always have i.i.d. samples from our target of interest. Data sets often exhibit biases—for example, under-representation of certain demographics—and if we ignore this fact our machine learning algorithms will propagate these biases. Alternatively, it may be useful to assume our data has been gathered via a biased sample selection mechanism in order to manipulate properties of the estimating distribution Q.
In this paper, we construct an estimator for the MMD between P and Q when we only have access to P via some biased sample selection mechanism, and suggest methods for estimating this sample selection mechanism when it is not already known. We show that this estimator can be used to train generative neural networ ks on a biased data sample, to give a simulator that reverses the effect of that bias.
****
Adversary A3C for Robust Reinforcement Learning
Zhaoyuan Gu,Zhenzhong Jia,Howie Choset
Asynchronous Advantage Actor Critic (A3C) is an effective Reinforcement Learning (RL) algorithm for a wide range of tasks, such as Atari games and robot control. The agent learns policies and value function through trial-and-error interacti ons with the environment until converging to an optimal policy. Robustness and s tability are critical in RL; however, neural network can be vulnerable to noise from unexpected sources and is not likely to withstand very slight disturbances. We note that agents generated from mild environment using A3C are not able to h andle challenging environments. Learning from adversarial examples, we proposed an algorithm called Adversary Robust A3C (AR-A3C) to improve the agent's perform ance under noisy environments. In this algorithm, an adversarial agent is introd uced to the learning process to make it more robust against adversarial disturba nces, thereby making it more adaptive to noisy environments. Both simulations an d real-world experiments are carried out to illustrate the stability of the prop osed algorithm. The AR-A3C algorithm outperforms A3C in both clean and noisy env ironments.
****
Faster Reinforcement Learning with Expert State Sequences
Xiaoxiao Guo,Shiyu Chang,Mo Yu,Miao Liu,Gerald Tesauro
Imitation learning relies on expert demonstrations. Existing approaches often re - quire that the complete demonstration data, including sequences of actions and states are available. In this paper, we consider a realistic and more difficult sce- nario where a reinforcement learning agent only has access to the state se quences of an expert, while the expert actions are not available. Inferring the unseen ex- pert actions in a stochastic environment is challenging and usually i nfeasible when combined with a large state space. We propose a novel policy lear ning method which only utilizes the expert state sequences without inferring the unseen ac- tions. Specifically, our agent first learns to extract useful sub-go al information from the state sequences of the expert and then utilizes the extr acted sub-goal information to factorize the action value estimate over state-act ion pairs and sub- goals. The extracted sub-goals are also used to synthesize gu idance rewards in the policy learning. We evaluate our agent on five Doom tasks. Our empirical results show that the proposed method significantly outperforms t he conventional DQN method.
****
Cross-View Training for Semi-Supervised Learning
Kevin Clark,Thang Luong,Quoc V. Le
We present Cross-View Training (CVT), a simple but effective method for deep sem i-supervised learning. On labeled examples, the model is trained with standard c ross-entropy loss. On an unlabeled example, the model first performs inference (acting as a "teacher") to produce soft targets. The model then learns from these

soft targets (acting as a ``"student"). We deviate from prior work by adding multiple auxiliary student prediction layers to the model. The input to each student layer is a sub-network of the full model that has a restricted view of the input (e.g., only seeing one region of an image). The students can learn from the teacher (the full model) because the teacher sees more of each example. Concurrently, the students improve the quality of the representations used by the teacher as they learn to make predictions with limited data. When combined with Virtual Adversarial Training, CVT improves upon the current state-of-the-art on semi-supervised CIFAR-10 and semi-supervised SVHN. We also apply CVT to train models on five natural language processing tasks using hundreds of millions of sentences of unlabeled data. On all tasks CVT substantially outperforms supervised learning alone, resulting in models that improve upon or are competitive with the current state-of-the-art.

****

Structured Deep Factorization Machine: Towards General-Purpose Architectures

José P. González-Brenes,Ralph Edezhath

In spite of their great success, traditional factorization algorithms typically do not support features (e.g., Matrix Factorization), or their complexity scales quadratically with the number of features (e.g, Factorization Machine). On the other hand, neural methods allow large feature sets, but are often designed for a specific application. We propose novel deep factorization methods that allow efficient and flexible feature representation. For example, we enable describing items with natural language with complexity linear to the vocabulary size—this enables prediction for unseen items and avoids the cold start problem. We show that our architecture can generalize some previously published single-purpose neural architectures. Our experiments suggest improved training times and accuracy compared to shallow methods.

****

Parameter Space Noise for Exploration

Matthias Plappert,Rein Houthooft,Prafulla Dhariwal,Szymon Sidor,Richard Y. Chen,Xi Chen,Tamim Asfour,Pieter Abbeel,Marcin Andrychowicz

Deep reinforcement learning (RL) methods generally engage in exploratory behavior through noise injection in the action space. An alternative is to add noise directly to the agent's parameters, which can lead to more consistent exploration and a richer set of behaviors. Methods such as evolutionary strategies use parameter perturbations, but discard all temporal structure in the process and require significantly more samples. Combining parameter noise with traditional RL methods allows to combine the best of both worlds. We demonstrate that both off- and on-policy methods benefit from this approach through experimental comparison of DQN, DDPG, and TRPO on high-dimensional discrete action environments as well as continuous control tasks.

****

ANALYSIS ON GRADIENT PROPAGATION IN BATCH NORMALIZED RESIDUAL NETWORKS

Abhishek Panigrahi,Yueru Chen,C.-C. Jay Kuo

We conduct a mathematical analysis on the Batch normalization (BN) effect on gradient backpropagation in residual network training in this work, which is believed to play a critical role in addressing the gradient vanishing/explosion problem. Specifically, by analyzing the mean and variance behavior of the input and the gradient in the forward and backward passes through the BN and residual branches, respectively, we show that they work together to confine the gradient variance to a certain range across residual blocks in backpropagation. As a result, the gradient vanishing/explosion problem is avoided. Furthermore, we use the same analysis to discuss the tradeoff between depth and width of a residual network and demonstrate that shallower yet wider resnets have stronger learning performance than deeper yet thinner resnets.

****

Compositional Attention Networks for Machine Reasoning

Drew A. Hudson,Christopher D. Manning

We present Compositional Attention Networks, a novel fully differentiable neural

network architecture, designed to facilitate explicit and expressive reasoning. While many types of neural networks are effective at learning and generalizing from massive quantities of data, this model moves away from monolithic black-box architectures towards a design that provides a strong prior for iterative reasoning, enabling it to support explainable and structured learning, as well as generalization from a modest amount of data. The model builds on the great success of existing recurrent cells such as LSTMs: It sequences a single recurrent Memory, Attention, and Control (MAC) cell, and by careful design imposes structural constraints on the operation of each cell and the interactions between them, incorporating explicit control and soft attention mechanisms into their interfaces. We demonstrate the model's strength and robustness on the challenging CLEVR data set for visual reasoning, achieving a new state-of-the-art 98.9% accuracy, halving the error rate of the previous best model. More importantly, we show that the new model is more computationally efficient, data-efficient, and requires an order of magnitude less time and/or data to achieve good results.
****

A novel method to determine the number of latent dimensions with SVD
Asana Neishabouri,Michel Desmarais
Determining the number of latent dimensions is a ubiquitous problem in machine learning. In this study, we introduce a novel method that relies on SVD to discover the number of latent dimensions. The general principle behind the method is to compare the curve of singular values of the SVD decomposition of a data set with the randomized data set curve. The inferred number of latent dimensions corresponds to the crossing point of the two curves. To evaluate our methodology, we compare it with competing methods such as Kaisers eigenvalue-greater-than-one rule (K1), Parallel Analysis (PA), Velicers MAP test (Minimum Average Partial). We also compare our method with the Silhouette Width (SW) technique which is used in different clustering methods to determine the optimal number of clusters. The result on synthetic data shows that the Parallel Analysis and our method have similar results and more accurate than the other methods, and that our methods is slightly better result than the Parallel Analysis method for the sparse data sets.
****

Lifelong Generative Modeling
Jason Ramapuram,Magda Gregorova,Alexandros Kalousis
Lifelong learning is the problem of learning multiple consecutive tasks in a sequential manner where knowledge gained from previous tasks is retained and used for future learning. It is essential towards the development of intelligent machines that can adapt to their surroundings. In this work we focus on a lifelong learning approach to generative modeling where we continuously incorporate newly observed streaming distributions into our learnt model. We do so through a student-teacher architecture which allows us to learn and preserve all the distributions seen so far without the need to retain the past data nor the past models. Through the introduction of a novel cross-model regularizer, the student model leverages the information learnt by the teacher, which acts as a summary of everything seen till now. The regularizer has the additional benefit of reducing the effect of catastrophic interference that appears when we learn over streaming data. We demonstrate its efficacy on streaming distributions as well as its ability to learn a common latent representation across a complex transfer learning scenario.

****

Bias-Variance Decomposition for Boltzmann Machines
Mahito Sugiyama,Koji Tsuda,Hiroyuki Nakahara
We achieve bias-variance decomposition for Boltzmann machines using an informati

on geometric formulation. Our decomposition leads to an interesting phenomenon that the variance does not necessarily increase when more parameters are included in Boltzmann machines, while the bias always decreases. Our result gives a theoretical evidence of the generalization ability of deep learning architectures because it provides the possibility of increasing the representation power with avoiding the variance inflation.
****

A dynamic game approach to training robust deep policies
Olalekan Ogunmolu
We present a method for evaluating the sensitivity of deep reinforcement learning (RL) policies. We also formulate a zero-sum dynamic game for designing robust deep reinforcement learning policies. Our approach mitigates the brittleness of policies when agents are trained in a simulated environment and are later exposed to the real world where it is hazardous to employ RL policies. This framework for training deep RL policies involve a zero-sum dynamic game against an adversarial agent, where the goal is to drive the system dynamics to a saddle region. Using a variant of the guided policy search algorithm, our agent learns to adopt robust policies that require less samples for learning the dynamics and performs better than the GPS algorithm. Without loss of generality, we demonstrate that deep RL policies trained in this fashion will be maximally robust to a ``worst" possible adversarial disturbances.
****

Generalizing Across Domains via Cross-Gradient Training
Shiv Shankar*,Vihari Piratla*,Soumen Chakrabarti,Siddhartha Chaudhuri,Preethi Jyothi,Sunita Sarawagi
We present CROSSGRAD , a method to use multi-domain training data to learn a classifier that generalizes to new domains. CROSSGRAD does not need an adaptation phase via labeled or unlabeled data, or domain features in the new domain. Most existing domain adaptation methods attempt to erase domain signals using techniques like domain adversarial training. In contrast, CROSSGRAD is free to use domain signals for predicting labels, if it can prevent overfitting on training domains. We conceptualize the task in a Bayesian setting, in which a sampling step is implemented as data augmentation, based on domain-guided perturbations of input instances. CROSSGRAD jointly trains a label and a domain classifier on examples perturbed by loss gradients of each other’s objectives. This enables us to directly perturb inputs, without separating and re-mixing domain signals while making various distributional assumptions. Empirical evaluation on three different applications where this setting is natural establishes that
(1) domain-guided perturbation provides consistently better generalization to unseen domains, compared to generic instance perturbation methods, and
(2) data augmentation is a more stable and accurate method than domain adversarial training.
****

Hierarchical Subtask Discovery with Non-Negative Matrix Factorization
Adam C. Earle,Andrew M. Saxe,Benjamin Rosman
Hierarchical reinforcement learning methods offer a powerful means of planning flexible behavior in complicated domains. However, learning an appropriate hierarchical decomposition of a domain into subtasks remains a substantial challenge. We present a novel algorithm for subtask discovery, based on the recently introduced multitask linearly-solvable Markov decision process (MLMDP) framework. The MLMDP can perform never-before-seen tasks by representing them as a linear combination of a previously learned basis set of tasks. In this setting, the subtask discovery problem can naturally be posed as finding an optimal low-rank approximation of the set of tasks the agent will face in a domain. We use non-negative matrix factorization to discover this minimal basis set of tasks, and show that the technique learns intuitive decompositions in a variety of domains. Our method has several qualitatively desirable features: it is not limited to learning subtasks with single goal states, instead learning distributed patterns of preferred states; it learns qualitatively different hierarchical decompositions in the same domain depending on the ensemble of tasks the agent will face; and it may be

straightforwardly iterated to obtain deeper hierarchical decompositions.
****

Variational image compression with a scale hyperprior
Johannes Ballé,David Minnen,Saurabh Singh,Sung Jin Hwang,Nick Johnston
We describe an end-to-end trainable model for image compression based on variational autoencoders. The model incorporates a hyperprior to effectively capture spatial dependencies in the latent representation. This hyperprior relates to side information, a concept universal to virtually all modern image codecs, but largely unexplored in image compression using artificial neural networks (ANNs). Unlike existing autoencoder compression methods, our model trains a complex prior jointly with the underlying autoencoder. We demonstrate that this model leads to state-of-the-art image compression when measuring visual quality using the popular MS-SSIM index, and yields rate--distortion performance surpassing published ANN-based methods when evaluated using a more traditional metric based on squared error (PSNR). Furthermore, we provide a qualitative comparison of models trained for different distortion metrics.
****

To Prune, or Not to Prune: Exploring the Efficacy of Pruning for Model Compression
Michael H. Zhu,Suyog Gupta
Model pruning seeks to induce sparsity in a deep neural network's various connection matrices, thereby reducing the number of nonzero-valued parameters in the model. Recent reports (Han et al., 2015; Narang et al., 2017) prune deep networks at the cost of only a marginal loss in accuracy and achieve a sizable reduction in model size. This hints at the possibility that the baseline models in these experiments are perhaps severely over-parameterized at the outset and a viable alternative for model compression might be to simply reduce the number of hidden units while maintaining the model's dense connection structure, exposing a similar trade-off in model size and accuracy. We investigate these two distinct paths for model compression within the context of energy-efficient inference in resource-constrained environments and propose a new gradual pruning technique that is simple and straightforward to apply across a variety of models/datasets with minimal tuning and can be seamlessly incorporated within the training process. We compare the accuracy of large, but pruned models (large-sparse) and their smaller, but dense (small-dense) counterparts with identical memory footprint. Across a broad range of neural network architectures (deep CNNs, stacked LSTM, and seq2seq LSTM models), we find large-sparse models to consistently outperform small-dense models and achieve up to 10x reduction in number of non-zero parameters with minimal loss in accuracy.
****

FigureQA: An Annotated Figure Dataset for Visual Reasoning
Samira Ebrahimi Kahou,Adam Atkinson,Vincent Michalski,Ákos Kádár,Adam Trischler,Yoshua Bengio
We introduce FigureQA, a visual reasoning corpus of over one million question-answer pairs grounded in over 100,000 images. The images are synthetic, scientific-style figures from five classes: line plots, dot-line plots, vertical and horizontal bar graphs, and pie charts. We formulate our reasoning task by generating questions from 15 templates; questions concern various relationships between plot elements and examine characteristics like the maximum, the minimum, area-under-the-curve, smoothness, and intersection. To resolve, such questions often require reference to multiple plot elements and synthesis of information distributed spatially throughout a figure. To facilitate the training of machine learning systems, the corpus also includes side data that can be used to formulate auxiliary objectives. In particular, we provide the numerical data used to generate each figure as well as bounding-box annotations for all plot elements. We study the proposed visual reasoning task by training several models, including the recently proposed Relation Network as strong baseline. Preliminary results indicate that the task poses a significant machine learning challenge. We envision FigureQA as a first step towards developing models that can intuitively recognize patterns from visual representations of data.

****
Deep Active Learning for Named Entity Recognition
Yanyao Shen,Hyokun Yun,Zachary C. Lipton,Yakov Kronrod,Animashree Anandkumar
Deep learning has yielded state-of-the-art performance on many natural language processing tasks including named entity recognition (NER). However, this typically requires large amounts of labeled data. In this work, we demonstrate that the amount of labeled training data can be drastically reduced when deep learning is combined with active learning. While active learning is sample-efficient, it can be computationally expensive since it requires iterative retraining. To speed this up, we introduce a lightweight architecture for NER, viz., the CNN-CNN-LSTM model consisting of convolutional character and word encoders and a long short term memory (LSTM) tag decoder. The model achieves nearly state-of-the-art performance on standard datasets for the task while being computationally much more efficient than best performing models. We carry out incremental active learning, during the training process, and are able to nearly match state-of-the-art performance with just 25\% of the original training data.
****
A Compressed Sensing View of Unsupervised Text Embeddings, Bag-of-n-Grams, and LSTMs
Sanjeev Arora,Mikhail Khodak,Nikunj Saunshi,Kiran Vodrahalli
Low-dimensional vector embeddings, computed using LSTMs or simpler techniques, are a popular approach for capturing the “meaning” of text and a form of unsupervised learning useful for downstream tasks. However, their power is not theoretically understood. The current paper derives formal understanding by looking at the subcase of linear embedding schemes. Using the theory of compressed sensing we show that representations combining the constituent word vectors are essentially information-preserving linear measurements of Bag-of-n-Grams (BonG) representations of text. This leads to a new theoretical result about LSTMs: low-dimensional embeddings derived from a low-memory LSTM are provably at least as powerful on classification tasks, up to small error, as a linear classifier over BonG vectors, a result that extensive empirical work has thus far been unable to show. Our experiments support these theoretical findings and establish strong, simple, and unsupervised baselines on standard benchmarks that in some cases are state of the art among word-level methods. We also show a surprising new property of embeddings such as GloVe and word2vec: they form a good sensing matrix for text that is more efficient than random matrices, the standard sparse recovery tool, which may explain why they lead to better representations in practice.
****
VOCABULARY-INFORMED VISUAL FEATURE AUGMENTATION FOR ONE-SHOT LEARNING
jianqi ma,hangyu lin,yinda zhang,yanwei fu,xiangyang xue
A natural solution for one-shot learning is to augment training data to handle the data deficiency problem. However, directly augmenting in the image domain may not necessarily generate training data that sufficiently explore the intra-class space for one-shot classification. Inspired by the recent vocabulary-informed learning, we propose to generate synthetic training data with the guide of the semantic word space. Essentially, we train an auto-encoder as a bridge to enable the transformation between the image feature space and the semantic space. Besides directly augmenting image features, we transform the image features to semantic space using the encoder and perform the data augmentation. The decoder then synthesizes the image features for the augmented instances from the semantic space. Experiments on three datasets show that our data augmentation method effectively improves the performance of one-shot classification. An extensive study shows that data augmented from semantic space are complementary with those from the image space, and thus boost the classification accuracy dramatically. Source code and dataset will be available.
****
Many Paths to Equilibrium: GANs Do Not Need to Decrease a Divergence At Every Step
William Fedus*,Mihaela Rosca*,Balaji Lakshminarayanan,Andrew M. Dai,Shakir Mohamed,Ian Goodfellow

Generative adversarial networks (GANs) are a family of generative models that do not minimize a single training criterion. Unlike other generative models, the data distribution is learned via a game between a generator (the generative model) and a discriminator (a teacher providing training signal) that each minimize their own cost. GANs are designed to reach a Nash equilibrium at which each player cannot reduce their cost without changing the other players' parameters. One useful approach for the theory of GANs is to show that a divergence between the training distribution and the model distribution obtains its minimum value at equilibrium. Several recent research directions have been motivated by the idea that this divergence is the primary guide for the learning process and that every step of learning should decrease the divergence. We show that this view is overly restrictive. During GAN training, the discriminator provides learning signal in situations where the gradients of the divergences between distributions would not be useful. We provide empirical counterexamples to the view of GAN training as divergence minimization. Specifically, we demonstrate that GANs are able to learn distributions in situations where the divergence minimization point of view predicts they would fail. We also show that gradient penalties motivated from the divergence minimization perspective are equally helpful when applied in other contexts in which the divergence minimization perspective does not predict they would be helpful. This contributes to a growing body of evidence that GAN training may be more usefully viewed as approaching Nash equilibria via trajectories that do not necessarily minimize a specific divergence at each step.
****
Multi-Mention Learning for Reading Comprehension with Neural Cascades
Swabha Swayamdipta,Ankur P. Parikh,Tom Kwiatkowski
Reading comprehension is a challenging task, especially when executed across longer or across multiple evidence documents, where the answer is likely to reoccur. Existing neural architectures typically do not scale to the entire evidence, and hence, resort to selecting a single passage in the document (either via truncation or other means), and carefully searching for the answer within that passage. However, in some cases, this strategy can be suboptimal, since by focusing on a specific passage, it becomes difficult to leverage multiple mentions of the same answer throughout the document. In this work, we take a different approach by constructing lightweight models that are combined in a cascade to find the answer. Each submodel consists only of feed-forward networks equipped with an attention mechanism, making it trivially parallelizable. We show that our approach can scale to approximately an order of magnitude larger evidence documents and can aggregate information from multiple mentions of each answer candidate across the document. Empirically, our approach achieves state-of-the-art performance on both the Wikipedia and web domains of the TriviaQA dataset, outperforming more complex, recurrent architectures.
****
Skip RNN: Learning to Skip State Updates in Recurrent Neural Networks
Víctor Campos,Brendan Jou,Xavier Giró-i-Nieto,Jordi Torres,Shih-Fu Chang
Recurrent Neural Networks (RNNs) continue to show outstanding performance in sequence modeling tasks. However, training RNNs on long sequences often face challenges like slow inference, vanishing gradients and difficulty in capturing long term dependencies. In backpropagation through time settings, these issues are tightly coupled with the large, sequential computational graph resulting from unfolding the RNN in time. We introduce the Skip RNN model which extends existing RNN models by learning to skip state updates and shortens the effective size of the computational graph. This model can also be encouraged to perform fewer state updates through a budget constraint. We evaluate the proposed model on various tasks and show how it can reduce the number of required RNN updates while preserving, and sometimes even improving, the performance of the baseline RNN models. Source code is publicly available at https://imatge-upc.github.io/skiprnn-2017-telecombcn/.
****
Modifying memories in a Recurrent Neural Network Unit
Vlad Velici,Adam Prügel-Bennett

Long Short-Term Memory (LSTM) units have the ability to memorise and use long-te
rm dependencies between inputs to generate predictions on time series data. We i
ntroduce the concept of modifying the cell state (memory) of LSTMs using rotatio
n matrices parametrised by a new set of trainable weights. This addition shows s
ignificant increases of performance on some of the tasks from the bAbI dataset.
****
Dynamic Neural Program Embeddings for Program Repair
Ke Wang,Rishabh Singh,Zhendong Su
Neural program embeddings have shown much promise recently for a variety of prog
ram analysis tasks, including program synthesis, program repair, code completion
, and fault localization. However, most existing program embeddings are based on
 syntactic features of programs, such as token sequences or abstract syntax tree
s. Unlike images and text, a program has well-de■ned semantics that can be dif■c
ult to capture by only considering its syntax (i.e. syntactically similar progra
ms can exhibit vastly different run-time behavior), which makes syntax-based pro
gram embeddings fundamentally limited. We propose a novel semantic program embed
ding that is learned from program execution traces. Our key insight is that prog
ram states expressed as sequential tuples of live variable values not only captu
re program semantics more precisely, but also offer a more natural ■t for Recurr
ent Neural Networks to model. We evaluate different syntactic and semantic progr
am embeddings on the task of classifying the types of errors that students make
in their submissions to an introductory programming class and on the CodeHunt ed
ucation platform. Our evaluation results show that the semantic program embeddin
gs signi■cantly outperform the syntactic program embeddings based on token seque
nces and abstract syntax trees. In addition, we augment a search-based program r
epair system with predictions made from our semantic embedding and demonstrate s
igni■cantly improved search ef■ciency.

****
Multiple Source Domain Adaptation with Adversarial Learning
Han Zhao,Shanghang Zhang,Guanhang Wu,Jo\~{a}o P. Costeira,Jos\'{e} M. F. Moura
,Geoffrey J. Gordon
While domain adaptation has been actively researched in recent years, most theor
etical results and algorithms focus on the single-source-single-target adaptatio
n setting. Naive application of such algorithms on multiple source domain adapta
tion problem may lead to suboptimal solutions. We propose a new generalization b
ound for domain adaptation when there are multiple source domains with labeled i
nstances and one target domain with unlabeled instances. Compared with existing
bounds, the new bound does not require expert knowledge about the target distrib
ution, nor the optimal combination rule for multisource domains. Interestingly,
our theory also leads to an efficient learning strategy using adversarial neural
 networks: we show how to interpret it as learning feature representations that
are invariant to the multiple domain shifts while still being discriminative for
 the learning task. To this end, we propose two models, both of which we call mu
ltisource domain adversarial networks (MDANs): the first model optimizes directl
y our bound, while the second model is a smoothed approximation of the first one
, leading to a more data-efficient and task-adaptive model. The optimization tas
ks of both models are minimax saddle point problems that can be optimized by adv
ersarial training. To demonstrate the effectiveness of MDANs, we conduct extensi
ve experiments showing superior adaptation performance on three real-world datas
ets: sentiment analysis, digit classification, and vehicle counting.

****
Can recurrent neural networks warp time?
Corentin Tallec,Yann Ollivier
Successful recurrent models such as long short-term memories (LSTMs) and gated r
ecurrent units (GRUs) use \emph{ad hoc} gating mechanisms. Empirically these mo
dels have been found to improve the learning of medium to long term temporal dep
endencies and to help with vanishing gradient issues.
■

We prove that learnable gates in a recurrent model formally provide \emph{quasi-
invariance to general time transformations} in the input data. We recover part o
f the LSTM architecture from a simple axiomatic approach.

This result leads to a new way of initializing gate biases in LSTMs and GRUs. Ex
perimentally, this new \emph{chrono initialization} is shown to greatly improve
learning of long term dependencies, with minimal implementation effort.

****
Consequentialist conditional cooperation in social dilemmas with imperfect infor
mation
Alexander Peysakhovich,Adam Lerer
Social dilemmas, where mutual cooperation can lead to high payoffs but participa
nts face incentives to cheat, are ubiquitous in multi-agent interaction. We wish
 to construct agents that cooperate with pure cooperators, avoid exploitation by
 pure defectors, and incentivize cooperation from the rest. However, often the a
ctions taken by a partner are (partially) unobserved or the consequences of indi
vidual actions are hard to predict. We show that in a large class of games good
strategies can be constructed by conditioning one's behavior solely on outcomes
(ie. one's past rewards). We call this consequentialist conditional cooperation.
 We show how to construct such strategies using deep reinforcement learning tech
niques and demonstrate, both analytically and experimentally, that they are effe
ctive in social dilemmas beyond simple matrix games. We also show the limitation
s of relying purely on consequences and discuss the need for understanding both
the consequences of and the intentions behind an action.
****
A Tensor Analysis on Dense Connectivity via Convolutional Arithmetic Circuits
Emilio Rafael Balda,Arash Behboodi,Rudolf Mathar
Several state of the art convolutional networks rely on inter-connecting differe
nt layers to ease the flow of information and gradient between their input and o
utput layers. These techniques have enabled practitioners to successfully train
deep convolutional networks with hundreds of layers. Particularly, a novel way o
f interconnecting layers was introduced as the Dense Convolutional Network (Dens
eNet) and has achieved state of the art performance on relevant image recognitio
n tasks. Despite their notable empirical success, their theoretical understandin
g is still limited. In this work, we address this problem by analyzing the effec
t of layer interconnection on the overall expressive power of a convolutional ne
twork. In particular, the connections used in DenseNet are compared with other
types of inter-layer connectivity. We carry out a tensor analysis on the express
ive power inter-connections on convolutional arithmetic circuits (ConvACs) and r
elate our results to standard convolutional networks. The analysis leads to perf
ormance bounds and practical guidelines for design of ConvACs. The generalizatio
n of these results are discussed for other kinds of convolutional networks via g
eneralized tensor decompositions.
****
CAYLEYNETS: SPECTRAL GRAPH CNNS WITH COMPLEX RATIONAL FILTERS
Ron Levie,Federico Monti,Xavier Bresson,Michael M. Bronstein
The rise of graph-structured data such as social networks, regulatory networks,
citation graphs, and functional brain networks, in combination with resounding s
uccess of deep learning in various applications, has brought the interest in gen
eralizing deep learning models to non-Euclidean domains.
In this paper, we introduce a new spectral domain convolutional architecture for
 deep learning on graphs. The core ingredient of our model is a new class of par
ametric rational complex functions (Cayley polynomials) allowing to efficiently
compute spectral filters on graphs that specialize on frequency bands of interes
t. Our model generates rich spectral filters that are localized in space, scales
 linearly with the size of the input data for sparsely-connected graphs, and can
 handle different constructions of Laplacian operators. Extensive experimental r
esults show the superior performance of our approach on spectral image classific

ation, community detection, vertex classification and matrix completion tasks.
****

Exploring the Space of Black-box Attacks on Deep Neural Networks
Arjun Nitin Bhagoji,Warren He,Bo Li,Dawn Song
Existing black-box attacks on deep neural networks (DNNs) so far have largely focused on transferability, where an adversarial instance generated for a locally trained model can “transfer” to attack other learning models. In this paper, we propose novel Gradient Estimation black-box attacks for adversaries with query access to the target model’s class probabilities, which do not rely on transferability. We also propose strategies to decouple the number of queries required to generate each adversarial sample from the dimensionality of the input. An iterative variant of our attack achieves close to 100% adversarial success rates for both targeted and untargeted attacks on DNNs. We carry out extensive experiments for a thorough comparative evaluation of black-box attacks and show that the proposed Gradient Estimation attacks outperform all transferability based black-box attacks we tested on both MNIST and CIFAR-10 datasets, achieving adversarial success rates similar to well known, state-of-the-art white-box attacks. We also apply the Gradient Estimation attacks successfully against a real-world content moderation classi■er hosted by Clarifai. Furthermore, we evaluate black-box attacks against state-of-the-art defenses. We show that the Gradient Estimation attacks are very effective even against these defenses.
****

Deep Generative Dual Memory Network for Continual Learning
Nitin Kamra,Umang Gupta,Yan Liu
Despite advances in deep learning, artificial neural networks do not learn the same way as humans do. Today, neural networks can learn multiple tasks when trained on them jointly, but cannot maintain performance on learnt tasks when tasks are presented one at a time -- this phenomenon called catastrophic forgetting is a fundamental challenge to overcome before neural networks can learn continually from incoming data. In this work, we derive inspiration from human memory to develop an architecture capable of learning continuously from sequentially incoming tasks, while averting catastrophic forgetting. Specifically, our model consists of a dual memory architecture to emulate the complementary learning systems (hippocampus and the neocortex) in the human brain and maintains a consolidated long-term memory via generative replay of past experiences. We (i) substantiate our claim that replay should be generative, (ii) show the benefits of generative replay and dual memory via experiments, and (iii) demonstrate improved performance retention even for small models with low capacity. Our architecture displays many important characteristics of the human memory and provides insights on the connection between sleep and learning in humans.
****

Dynamic Evaluation of Neural Sequence Models
Ben Krause,Emmanuel Kahembwe,Iain Murray,Steve Renals
We present methodology for using dynamic evaluation to improve neural sequence models. Models are adapted to recent history via a gradient descent based mechanism, causing them to assign higher probabilities to re-occurring sequential patterns. Dynamic evaluation outperforms existing adaptation approaches in our comparisons. Dynamic evaluation improves the state-of-the-art word-level perplexities on the Penn Treebank and WikiText-2 datasets to 51.1 and 44.3 respectively, and the state-of-the-art character-level cross-entropies on the text8 and Hutter Prize datasets to 1.19 bits/char and 1.08 bits/char respectively.
****

Lifelong Learning with Output Kernels
Keerthiram Murugesan,Jaime Carbonell
Lifelong learning poses considerable challenges in terms of effectiveness (minimizing prediction errors for all tasks) and overall computational tractability for real-time performance. This paper addresses continuous lifelong multitask learning by jointly re-estimating the inter-task relations (\textit{output} kernel) and the per-task model parameters at each round, assuming data arrives in a streaming fashion. We propose a novel algorithm called \textit{Online Output Kerne

l Learning Algorithm} (OOKLA) for lifelong learning setting. To avoid the memory explosion, we propose a robust budget-limited versions of the proposed algorith m that efficiently utilize the relationship between the tasks to bound the total number of representative examples in the support set. In addition, we propose a two-stage budgeted scheme for efficiently tackling the task-specific budget co nstraints in lifelong learning. Our empirical results over three datasets indica te superior AUC performance for OOKLA and its budget-limited cousins over strong baselines.
****
ResBinNet: Residual Binary Neural Network
Mohammad Ghasemzadeh,Mohammad Samragh,Farinaz Koushanfar
Recent efforts on training light-weight binary neural networks offer promising e xecution/memory efficiency. This paper introduces ResBinNet, which is a composit ion of two interlinked methodologies aiming to address the slow convergence spee d and limited accuracy of binary convolutional neural networks. The first method , called residual binarization, learns a multi-level binary representation for t he features within a certain neural network layer. The second method, called tem perature adjustment, gradually binarizes the weights of a particular layer. The two methods jointly learn a set of soft-binarized parameters that improve the co nvergence rate and accuracy of binary neural networks. We corroborate the applic ability and scalability of ResBinNet by implementing a prototype hardware accele rator. The accelerator is reconfigurable in terms of the numerical precision of the binarized features, offering a trade-off between runtime and inference accur acy.

****
On the Generalization Effects of DenseNet Model Structures
Yin Liu,Vincent Chen
Modern neural network architectures take advantage of increasingly deeper layers , and various advances in their structure to achieve better performance. While t raditional explicit regularization techniques like dropout, weight decay, and da ta augmentation are still being used in these new models, little about the regul arization and generalization effects of these new structures have been studied. Besides being deeper than their predecessors, could newer architectures like Res Net and DenseNet also benefit from their structures' implicit regularization pro perties?
In this work, we investigate the skip connection's effect on network's generaliz ation features. Through experiments, we show that certain neural network archite ctures contribute to their generalization abilities. Specifically, we study the effect that low-level features have on generalization performance when they are introduced to deeper layers in DenseNet, ResNet as well as networks with 'skip c onnections'. We show that these low-level representations do help with generaliz ation in multiple settings when both the quality and quantity of training data i s decreased.
****
FAST READING COMPREHENSION WITH CONVNETS
Felix Wu,Ni Lao,John Blitzer,Guandao Yang,Kilian Weinberger
State-of-the-art deep reading comprehension models are dominated by recurrent neural nets. Their sequential nature is a natural fit for language, but it also precludes
parallelization within an instances and often becomes the bottleneck for deploying such models to latency critical scenarios. This is particularly proble matic
for longer texts. Here we present a convolutional architecture as an alternative to these recurrent architectures. Using simple dilated convolutional units in pl ace
of recurrent ones, we achieve results comparable to the state of the art on two question answering tasks, while at the same time achieving up to two orders of magnitude speedups for question answering.
****

Generating Natural Adversarial Examples
Zhengli Zhao,Dheeru Dua,Sameer Singh
Due to their complex nature, it is hard to characterize the ways in which machin e learning models can misbehave or be exploited when deployed. Recent work on ad versarial examples, i.e. inputs with minor perturbations that result in substant ially different model predictions, is helpful in evaluating the robustness of th ese models by exposing the adversarial scenarios where they fail. However, these malicious perturbations are often unnatural, not semantically meaningful, and n ot applicable to complicated domains such as language. In this paper, we propose a framework to generate natural and legible adversarial examples that lie on th e data manifold, by searching in semantic space of dense and continuous data rep resentation, utilizing the recent advances in generative adversarial networks. W e present generated adversaries to demonstrate the potential of the proposed app roach for black-box classifiers for a wide range of applications such as image c lassification, textual entailment, and machine translation. We include experimen ts to show that the generated adversaries are natural, legible to humans, and us eful in evaluating and analyzing black-box classifiers.
****
Large Scale Optimal Transport and Mapping Estimation
Vivien Seguy,Bharath Bhushan Damodaran,Remi Flamary,Nicolas Courty,Antoine Rolet ,Mathieu Blondel
This paper presents a novel two-step approach for the fundamental problem of lea rning an optimal map from one distribution to another. First, we learn an optima l transport (OT) plan, which can be thought as a one-to-many map between the two distributions. To that end, we propose a stochastic dual approach of regularize d OT, and show empirically that it scales better than a recent related approach when the amount of samples is very large. Second, we estimate a Monge map as a d eep neural network learned by approximating the barycentric projection of the pr eviously-obtained OT plan. This parameterization allows generalization of the ma pping outside the support of the input measure. We prove two theoretical stabili ty results of regularized OT which show that our estimations converge to the OT and Monge map between the underlying continuous measures. We showcase our propos ed approach on two applications: domain adaptation and generative modeling.
****
Learning to Generate Filters for Convolutional Neural Networks
Wei Shen,Rujie Liu
Conventionally, convolutional neural networks (CNNs) process different images wi th the same set of filters. However, the variations in images pose a challenge t o this fashion. In this paper, we propose to generate sample-specific filters fo r convolutional layers in the forward pass. Since the filters are generated on-t he-fly, the model becomes more flexible and can better fit the training data com pared to traditional CNNs. In order to obtain sample-specific features, we extra ct the intermediate feature maps from an autoencoder. As filters are usually hig h dimensional, we propose to learn a set of coefficients instead of a set of fil ters. These coefficients are used to linearly combine the base filters from a fi lter repository to generate the final filters for a CNN. The proposed method is evaluated on MNIST, MTFL and CIFAR10 datasets. Experiment results demonstrate th at the classification accuracy of the baseline model can be improved by using th e proposed filter generation method.
****
Structured Exploration via Hierarchical Variational Policy Networks
Stephan Zheng,Yisong Yue
Reinforcement learning in environments with large state-action spaces is challen ging, as exploration can be highly inefficient. Even if the dynamics are simple, the optimal policy can be combinatorially hard to discover. In this work, we pr opose a hierarchical approach to structured exploration to improve the sample ef ficiency of on-policy exploration in large state-action spaces. The key idea is to model a stochastic policy as a hierarchical latent variable model, which can learn low-dimensional structure in the state-action space, and to define explora tion by sampling from the low-dimensional latent space. This approach enables lo

wer sample complexity, while preserving policy expressivity. In order to make learning tractable, we derive a joint learning and exploration strategy by combining hierarchical variational inference with actor-critic learning. The benefits of our learning approach are that 1) it is principled, 2) simple to implement, 3) easily scalable to settings with many actions and 4) easily composable with existing deep learning approaches. We demonstrate the effectiveness of our approach on learning a deep centralized multi-agent policy, as multi-agent environments naturally have an exponentially large state-action space. In this setting, the latent hierarchy implements a form of multi-agent coordination during exploration and execution (MACE). We demonstrate empirically that MACE can more efficiently learn optimal policies in challenging multi-agent games with a large number (~20) of agents, compared to conventional baselines. Moreover, we show that our hierarchical structure leads to meaningful agent coordination.
****

Label Embedding Network: Learning Label Representation for Soft Training of Deep Networks
Xu Sun,Bingzhen Wei,Xuancheng Ren,Shuming Ma
We propose a method, called Label Embedding Network, which can learn label representation (label embedding) during the training process of deep networks. With the proposed method, the label embedding is adaptively and automatically learned through back propagation. The original one-hot represented loss function is converted into a new loss function with soft distributions, such that the originally unrelated labels have continuous interactions with each other during the training process. As a result, the trained model can achieve substantially higher accuracy and with faster convergence speed. Experimental results based on competitive tasks demonstrate the effectiveness of the proposed method, and the learned label embedding is reasonable and interpretable. The proposed method achieves comparable or even better results than the state-of-the-art systems.
****

Taking Apart Autoencoders: How do They Encode Geometric Shapes ?
Alasdair Newson,Andres Almansa,Yann Gousseau,Said Ladjal
We study the precise mechanisms which allow autoencoders to encode and decode a simple geometric shape, the disk. In this carefully controlled setting, we are able to describe the specific form of the optimal solution to the minimisation problem of the training step. We show that the autoencoder indeed approximates this solution during training. Secondly, we identify a clear failure in the generalisation capacity of the autoencoder, namely its inability to interpolate data. Finally, we explore several regularisation schemes to resolve the generalisation problem. Given the great attention that has been recently given to the generative capacity of neural networks, we believe that studying in depth simple geometric cases sheds some light on the generation process and can provide a minimal requirement experimental setup for more complex architectures.

****

Stabilizing Adversarial Nets with Prediction Methods
Abhay Yadav,Sohil Shah,Zheng Xu,David Jacobs,Tom Goldstein
Adversarial neural networks solve many important problems in data science, but are notoriously difficult to train. These difficulties come from the fact that optimal weights for adversarial nets correspond to saddle points, and not minimizers, of the loss function. The alternating stochastic gradient methods typically used for such problems do not reliably converge to saddle points, and when convergence does happen it is often highly sensitive to learning rates. We propose a simple modification of stochastic gradient descent that stabilizes adversarial networks. We show, both in theory and practice, that the proposed method reliably converges to saddle points. This makes adversarial networks less likely to "collapse," and enables faster training with larger learning rates.
****

DeepArchitect: Automatically Designing and Training Deep Architectures
Renato Negrinho,Geoff Gordon
In deep learning, performance is strongly affected by the choice of architecture

and hyperparameters. While there has been extensive work on automatic hyperparameter optimization for simple spaces, complex spaces such as the space of deep architectures remain largely unexplored. As a result, the choice of architecture is
done manually by the human expert through a slow trial and error process guided mainly by intuition. In this paper we describe a framework for automatically designing and training deep models. We propose an extensible and modular language that allows the human expert to compactly represent complex search spaces over architectures and their hyperparameters. The resulting search spaces are tree-
structured and therefore easy to traverse. Models can be automatically compiled to
computational graphs once values for all hyperparameters have been chosen. We can leverage the structure of the search space to introduce different model search
algorithms, such as random search, Monte Carlo tree search (MCTS), and sequential model-based optimization (SMBO). We present experiments comparing the different algorithms on CIFAR-10 and show that MCTS and SMBO outperform random search. We also present experiments on MNIST, showing that the same search space achieves near state-of-the-art performance with a few samples. These
experiments show that our framework can be used effectively for model discovery, as it is possible to describe expressive search spaces and discover competitive
models without much effort from the human expert. Code for our framework and experiments has been made publicly available
****

SEARNN: Training RNNs with global-local losses
Rémi Leblond,Jean-Baptiste Alayrac,Anton Osokin,Simon Lacoste-Julien
We propose SEARNN, a novel training algorithm for recurrent neural networks (RNNs) inspired by the "learning to search" (L2S) approach to structured prediction. RNNs have been widely successful in structured prediction applications such as machine translation or parsing, and are commonly trained using maximum likelihood estimation (MLE). Unfortunately, this training loss is not always an appropriate surrogate for the test error: by only maximizing the ground truth probability, it fails to exploit the wealth of information offered by structured losses. Further, it introduces discrepancies between training and predicting (such as exposure bias) that may hurt test performance. Instead, SEARNN leverages test-alike search space exploration to introduce global-local losses that are closer to the test error. We first demonstrate improved performance over MLE on two different tasks: OCR and spelling correction. Then, we propose a subsampling strategy to enable SEARNN to scale to large vocabulary sizes. This allows us to validate the benefits of our approach on a machine translation task.
****

Enhance Word Representation for Out-of-Vocabulary on Ubuntu Dialogue Corpus
JIANXIONG DONG,Jim Huang
Ubuntu dialogue corpus is the largest public available dialogue corpus to make it feasible to build end-to-end
deep neural network models directly from the conversation data. One challenge of Ubuntu dialogue corpus is
the large number of out-of-vocabulary words. In this paper we proposed an algorithm which combines the general pre-trained word embedding vectors with those generated on the task-specific training set to address this issue. We integrated character embedding into Chen et al's Enhanced LSTM method (ESIM) and used it to evaluate the effectiveness of our proposed method. For the task of next utterance selection, the proposed method has demonstrated a significant performance improvement against original ESIM and the new model has achieved state-of-the-art results on both Ubuntu dialogue corpus and Douban conversation corpus. In addition, we investigated the performance impact of end-of-utterance and end-of-turn token tags.

****
Simulated+Unsupervised Learning With Adaptive Data Generation and Bidirectional Mappings
Kangwook Lee,Hoon Kim,Changho Suh
Collecting a large dataset with high quality annotations is expensive and time-consuming. Recently, Shrivastava et al. (2017) propose Simulated+Unsupervised (S+U) learning: It first learns a mapping from synthetic data to real data, translates a large amount of labeled synthetic data to the ones that resemble real data, and then trains a learning model on the translated data. Bousmalis et al. (2017) propose a similar framework that jointly trains a translation mapping and a learning model.
While these algorithms are shown to achieve the state-of-the-art performances on various tasks, it may have a room for improvement, as they do not fully leverage flexibility of data simulation process and consider only the forward (synthetic to real) mapping. While these algorithms are shown to achieve the state-of-the-art performances on various tasks, it may have a room for improvement, as it does not fully leverage flexibility of data simulation process and consider only the forward (synthetic to real) mapping. Inspired by this limitation, we propose a new S+U learning algorithm, which fully leverage the flexibility of data simulators and bidirectional mappings between synthetic data and real data. We show that our approach achieves the improved performance on the gaze estimation task, outperforming (Shrivastava et al., 2017).
****
Beyond Shared Hierarchies: Deep Multitask Learning through Soft Layer Ordering
Elliot Meyerson,Risto Miikkulainen
Existing deep multitask learning (MTL) approaches align layers shared between tasks in a parallel ordering. Such an organization significantly constricts the types of shared structure that can be learned. The necessity of parallel ordering for deep MTL is first tested by comparing it with permuted ordering of shared layers. The results indicate that a flexible ordering can enable more effective sharing, thus motivating the development of a soft ordering approach, which learns how shared layers are applied in different ways for different tasks. Deep MTL with soft ordering outperforms parallel ordering methods across a series of domains. These results suggest that the power of deep MTL comes from learning highly general building blocks that can be assembled to meet the demands of each task.
****
Sequential Coordination of Deep Models for Learning Visual Arithmetic
Eric Crawford,Guillaume Rabusseau,Joelle Pineau
Achieving machine intelligence requires a smooth integration of perception and reasoning, yet models developed to date tend to specialize in one or the other; sophisticated manipulation of symbols acquired from rich perceptual spaces has so far proved elusive. Consider a visual arithmetic task, where the goal is to carry out simple arithmetical algorithms on digits presented under natural conditions (e.g. hand-written, placed randomly). We propose a two-tiered architecture for tackling this kind of problem. The lower tier consists of a heterogeneous collection of information processing modules, which can include pre-trained deep neural networks for locating and extracting characters from the image, as well as modules performing symbolic transformations on the representations extracted by perception. The higher tier consists of a controller, trained using reinforcement learning, which coordinates the modules in order to solve the high-level task. For instance, the controller may learn in what contexts to execute the perceptual networks and what symbolic transformations to apply to their outputs. The resulting model is able to solve a variety of tasks in the visual arithmetic domain, and has several advantages over standard, architecturally homogeneous feedforward networks including improved sample efficiency.
****
Towards better understanding of gradient-based attribution methods for Deep Neural Networks
Marco Ancona,Enea Ceolini,Cengiz Öztireli,Markus Gross
Understanding the flow of information in Deep Neural Networks (DNNs) is a challe

nging problem that has gain increasing attention over the last few years. While several methods have been proposed to explain network predictions, there have been only a few attempts to compare them from a theoretical perspective. What is more, no exhaustive empirical comparison has been performed in the past. In this work we analyze four gradient-based attribution methods and formally prove conditions of equivalence and approximation between them. By reformulating two of these methods, we construct a unified framework which enables a direct comparison, as well as an easier implementation. Finally, we propose a novel evaluation metric, called Sensitivity-n and test the gradient-based attribution methods alongside with a simple perturbation-based attribution method on several datasets in the domains of image and text classification, using various network architectures.
****

Unsupervised Machine Translation Using Monolingual Corpora Only

Guillaume Lample,Alexis Conneau,Ludovic Denoyer,Marc'Aurelio Ranzato

Machine translation has recently achieved impressive performance thanks to recent advances in deep learning and the availability of large-scale parallel corpora. There have been numerous attempts to extend these successes to low-resource language pairs, yet requiring tens of thousands of parallel sentences. In this work, we take this research direction to the extreme and investigate whether it is possible to learn to translate even without any parallel data. We propose a model that takes sentences from monolingual corpora in two different languages and maps them into the same latent space. By learning to reconstruct in both languages from this shared feature space, the model effectively learns to translate without using any labeled data. We demonstrate our model on two widely used datasets and two language pairs, reporting BLEU scores of 32.8 and 15.1 on the Multi30k and WMT English-French datasets, without using even a single parallel sentence at training time.
****

Forced Apart: Discovering Disentangled Representations Without Exhaustive Labels

Alexey Romanov,Anna Rumshisky

Learning a better representation with neural networks is a challenging problem, which has been tackled from different perspectives in the past few years. In this work, we focus on learning a representation that would be useful in a clustering task. We introduce two novel loss components that substantially improve the quality of produced clusters, are simple to apply to arbitrary models and cost functions, and do not require a complicated training procedure. We perform an extensive set of experiments, supervised and unsupervised, and evaluate the proposed loss components on two most common types of models, Recurrent Neural Networks and Convolutional Neural Networks, showing that the approach we propose consistently improves the quality of KMeans clustering in terms of mutual information scores and outperforms previously proposed methods.
****

NerveNet: Learning Structured Policy with Graph Neural Networks

Tingwu Wang,Renjie Liao,Jimmy Ba,Sanja Fidler

We address the problem of learning structured policies for continuous control. In traditional reinforcement learning, policies of agents are learned by MLPs which take the concatenation of all observations from the environment as input for predicting actions. In this work, we propose NerveNet to explicitly model the structure of an agent, which naturally takes the form of a graph. Specifically, serving as the agent's policy network, NerveNet first propagates information over the structure of the agent and then predict actions for different parts of the agent. In the experiments, we first show that our NerveNet is comparable to state-of-the-art methods on standard MuJoCo environments. We further propose our customized reinforcement learning environments for benchmarking two types of structure transfer learning tasks, i.e., size and disability transfer. We demonstrate that policies learned by NerveNet are significantly better than policies learned by other models and are able to transfer even in a zero-shot setting.

****

Modeling Latent Attention Within Neural Networks

Christopher Grimm,Dilip Arumugam,Siddharth Karamcheti,David Abel,Lawson L.S. Wong,Michael L. Littman
Deep neural networks are able to solve tasks across a variety of domains and modalities of data. Despite many empirical successes, we lack the ability to clearly understand and interpret the learned mechanisms that contribute to such effective behaviors and more critically, failure modes. In this work, we present a general method for visualizing an arbitrary neural network's inner mechanisms and their power and limitations. Our dataset-centric method produces visualizations of how a trained network attends to components of its inputs. The computed "attention masks" support improved interpretability by highlighting which input attributes are critical in determining output. We demonstrate the effectiveness of our framework on a variety of deep neural network architectures in domains from computer vision and natural language processing. The primary contribution of our approach is an interpretable visualization of attention that provides unique insights into the network's underlying decision-making process irrespective of the data modality.
****

Training Deep AutoEncoders for Recommender Systems

Oleksii Kuchaiev,Boris Ginsburg
This paper proposes a new model for the rating prediction task in recommender systems which significantly outperforms previous state-of-the art models on a time-split Netflix data set. Our model is based on deep autoencoder with 6 layers and is trained end-to-end without any layer-wise pre-training. We empirically demonstrate that: a) deep autoencoder models generalize much better than the shallow ones, b) non-linear activation functions with negative parts are crucial for training deep models, and c) heavy use of regularization techniques such as dropout is necessary to prevent over-fitting. We also propose a new training algorithm based on iterative output re-feeding to overcome natural sparseness of collaborate filtering. The new algorithm significantly speeds up training and improves model performance. Our code is publicly available.
****

Predicting Auction Price of Vehicle License Plate with Deep Recurrent Neural Network

Vinci Chow
In Chinese societies, superstition is of paramount importance, and vehicle license plates with desirable numbers can fetch very high prices in auctions. Unlike other valuable items, license plates are not allocated an estimated price before auction.

I propose that the task of predicting plate prices can be viewed as a natural language processing (NLP) task, as the value depends on the meaning of each individual character on the plate and its semantics. I construct a deep recurrent neural network (RNN) to predict the prices of vehicle license plates in Hong Kong, based on the characters on a plate. I demonstrate the importance of having a deep network and of retraining. Evaluated on 13 years of historical auction prices, the deep RNN's predictions can explain over 80 percent of price variations, outperforming previous models by a significant margin. I also demonstrate how the model can be extended to become a search engine for plates and to provide estimates of the expected price distribution.
****

Leave no Trace: Learning to Reset for Safe and Autonomous Reinforcement Learning

Benjamin Eysenbach,Shixiang Gu,Julian Ibarz,Sergey Levine
Deep reinforcement learning algorithms can learn complex behavioral skills, but real-world application of these methods requires a considerable amount of experience to be collected by the agent. In practical settings, such as robotics, this involves repeatedly attempting a task, resetting the environment between each attempt. However, not all tasks are easily or automatically reversible. In practice, this learning process requires considerable human intervention. In this work, we propose an autonomous method for safe and efficient reinforcement learning that simultaneously learns a forward and backward policy, with the backward poli

cy resetting the environment for a subsequent attempt. By learning a value funct
ion for the backward policy, we can automatically determine when the forward pol
icy is about to enter a non-reversible state, providing for uncertainty-aware sa
fety aborts. Our experiments illustrate that proper use of the backward policy c
an greatly reduce the number of manual resets required to learn a task and can r
educe the number of unsafe actions that lead to non-reversible states.
****
Neural Language Modeling by Jointly Learning Syntax and Lexicon
Yikang Shen,Zhouhan Lin,Chin-wei Huang,Aaron Courville
We propose a neural language model capable of unsupervised syntactic structure i
nduction. The model leverages the structure information to form better semantic
representations and better language modeling. Standard recurrent neural networks
 are limited by their structure and fail to efficiently use syntactic informatio
n. On the other hand, tree-structured recursive networks usually require additio
nal structural supervision at the cost of human expert annotation. In this paper
, We propose a novel neural language model, called the Parsing-Reading-Predict N
etworks (PRPN), that can simultaneously induce the syntactic structure from unan
notated sentences and leverage the inferred structure to learn a better language
 model. In our model, the gradient can be directly back-propagated from the lang
uage model loss into the neural parsing network. Experiments show that the propo
sed model can discover the underlying syntactic structure and achieve state-of-t
he-art performance on word/character-level language model tasks.
****
WHAT ARE GANS USEFUL FOR?
Pablo M. Olmos,Briland Hitaj,Paolo Gasti,Giuseppe Ateniese,Fernando Perez-Cruz
GANs have shown how deep neural networks can be used for generative modeling, ai
ming at achieving the same impact that they brought for discriminative modeling.
 The first results were impressive, GANs were shown to be able to generate sampl
es in high dimensional structured spaces, like images and text, that were no cop
ies of the training data. But generative and discriminative learning are quite d
ifferent. Discriminative learning has a clear end, while generative modeling is
an intermediate step to understand the data or generate hypothesis. The quality
of implicit density estimation is hard to evaluate, because we cannot tell how w
ell a data is represented by the model. How can we certainly say that a generati
ve process is generating natural images with the same distribution as we do? In
this paper, we noticed that even though GANs might not be able to generate sampl
es from the underlying distribution (or we cannot tell at least), they are captu
ring some structure of the data in that high dimensional space. It is therefore
needed to address how we can leverage those estimates produced by GANs in the sa
me way we are able to use other generative modeling algorithms.
****
Compact Encoding of Words for Efficient Character-level Convolutional Neural Net
works Text Classification
Wemerson Marinho,Luis Marti,Nayat Sanchez-pi
This paper puts forward a new text to tensor representation that relies on infor
mation compression techniques to assign shorter codes to the most frequently use
d characters. This representation is language-independent with no need of pretra
ining and produces an encoding with no information loss. It provides an adequate
 description of the morphology of text, as it is able to represent prefixes, dec
lensions, and inflections with similar vectors and are able to represent even un
seen words on the training dataset. Similarly, as it is compact yet sparse, is i
deal for speed up training times using tensor processing libraries. As part of t
his paper, we show that this technique is especially effective when coupled with
 convolutional neural networks (CNNs) for text classification at character-level
. We apply two variants of CNN coupled with it. Experimental results show that i
t drastically reduces the number of parameters to be optimized, resulting in com
petitive classification accuracy values in only a fraction of the time spent by
one-hot encoding representations, thus enabling training in commodity hardware.
****
Deep Boosting of Diverse Experts

Wei Zhang,Qiuyu Chen,Jun Yu,Jianping Fan
In this paper, a deep boosting algorithm is developed to
learn more discriminative ensemble classifier by seamlessly combining a set of base deep CNNs (base experts)
with diverse capabilities, e.g., these base deep CNNs are
sequentially trained to recognize a set of
object classes in an easy-to-hard way according to their
learning complexities. Our experimental results have demonstrated
that our deep boosting algorithm can significantly improve the
accuracy rates on large-scale visual recognition.
****

Adaptive Dropout with Rademacher Complexity Regularization
Ke Zhai,Huan Wang
We propose a novel framework to adaptively adjust the dropout rates for the deep neural network based on a Rademacher complexity bound. The state-of-the-art deep learning algorithms impose dropout strategy to prevent feature co-adaptation. However, choosing the dropout rates remains an art of heuristics or relies on empirical grid-search over some hyperparameter space. In this work, we show the network Rademacher complexity is bounded by a function related to the dropout rate vectors and the weight coefficient matrices. Subsequently, we impose this bound as a regularizer and provide a theoretical justified way to trade-off between model complexity and representation power. Therefore, the dropout rates and the empirical loss are unified into the same objective function, which is then optimized using the block coordinate descent algorithm. We discover that the adaptively adjusted dropout rates converge to some interesting distributions that reveal meaningful patterns.Experiments on the task of image and document classification also show our method achieves better performance compared to the state-of the-art dropout algorithms.
****

Implicit Causal Models for Genome-wide Association Studies
Dustin Tran,David M. Blei
Progress in probabilistic generative models has accelerated, developing richer models with neural architectures, implicit densities, and with scalable algorithms for their Bayesian inference. However, there has been limited progress in models that capture causal relationships, for example, how individual genetic factors cause major human diseases. In this work, we focus on two challenges in particular: How do we build richer causal models, which can capture highly nonlinear relationships and interactions between multiple causes? How do we adjust for latent confounders, which are variables influencing both cause and effect and which prevent learning of causal relationships? To address these challenges, we synthesize ideas from causality and modern probabilistic modeling. For the first, we describe implicit causal models, a class of causal models that leverages neural architectures with an implicit density. For the second, we describe an implicit causal model that adjusts for confounders by sharing strength across examples. In experiments, we scale Bayesian inference on up to a billion genetic measurements. We achieve state of the art accuracy for identifying causal factors: we significantly outperform the second best result by an absolute difference of 15-45.3%.
****

Training GANs with Optimism
Constantinos Daskalakis,Andrew Ilyas,Vasilis Syrgkanis,Haoyang Zeng
We address the issue of limit cycling behavior in training Generative Adversarial Networks and propose the use of Optimistic Mirror Decent (OMD) for training Wasserstein GANs. Recent theoretical results have shown that optimistic mirror decent (OMD) can enjoy faster regret rates in the context of zero-sum games. WGANs is exactly a context of solving a zero-sum game with simultaneous no-regret dynamics. Moreover, we show that optimistic mirror decent addresses the limit cycling problem in training WGANs. We formally show that in the case of bi-linear zero-sum games the last iterate of OMD dynamics converges to an equilibrium, in contrast to GD dynamics which are bound to cycle. We also portray the huge qualitat

ive difference between GD and OMD dynamics with toy examples, even when GD is mo
dified with many adaptations proposed in the recent literature, such as gradient
 penalty or momentum. We apply OMD WGAN training to a bioinformatics problem of
generating DNA sequences. We observe that models trained with OMD achieve consis
tently smaller KL divergence with respect to the true underlying distribution, t
han models trained with GD variants. Finally, we introduce a new algorithm, Opti
mistic Adam, which is an optimistic variant of Adam. We apply it to WGAN trainin
g on CIFAR10 and observe improved performance in terms of inception score as com
pared to Adam.
****
Weightless: Lossy Weight Encoding For Deep Neural Network Compression
Brandon Reagen,Udit Gupta,Robert Adolf,Michael Mitzenmacher,Alexander Rush,Gu-Ye
on Wei,David Brooks
The large memory requirements of deep neural networks strain the capabilities of
 many devices, limiting their deployment and adoption. Model compression methods
 effectively reduce the memory requirements of these models, usually through app
lying transformations such as weight pruning or quantization. In this paper, we
present a novel scheme for lossy weight encoding which complements conventional
compression techniques. The encoding is based on the Bloomier filter, a probabil
istic data structure that can save space at the cost of introducing random error
s. Leveraging the ability of neural networks to tolerate these imperfections and
 by re-training around the errors, the proposed technique, Weightless, can compr
ess DNN weights by up to 496x; with the same model accuracy, this results in up
to a 1.51x improvement over the state-of-the-art.
****
Large-scale Cloze Test Dataset Designed by Teachers
Qizhe Xie,Guokun Lai,Zihang Dai,Eduard Hovy
Cloze test is widely adopted in language exams to evaluate students' language pr
oficiency. In this paper, we propose the first large-scale human-designed cloze
test dataset CLOTH in which the questions were used in middle-school and high-sc
hool language exams. With the missing blanks carefully created by teachers and c
andidate choices purposely designed to be confusing, CLOTH requires a deeper lan
guage understanding and a wider attention span than previous automatically gener
ated cloze datasets. We show humans outperform dedicated designed baseline model
s by a significant margin, even when the model is trained on sufficiently large
external data. We investigate the source of the performance gap, trace model def
iciencies to some distinct properties of CLOTH, and identify the limited ability
 of comprehending a long-term context to be the key bottleneck. In addition, we
find that human-designed data leads to a larger gap between the model's performa
nce and human performance when compared to automatically generated data.
****
Accelerating Neural Architecture Search using Performance Prediction
Bowen Baker*,Otkrist Gupta*,Ramesh Raskar,Nikhil Naik
Methods for neural network hyperparameter optimization and meta-modeling are com
putationally expensive due to the need to train a large number of model configur
ations. In this paper, we show that standard frequentist regression models can p
redict the final performance of partially trained model configurations using fea
tures based on network architectures, hyperparameters, and time series validatio
n performance data. We empirically show that our performance prediction models a
re much more effective than prominent Bayesian counterparts, are simpler to impl
ement, and are faster to train. Our models can predict final performance in both
 visual classification and language modeling domains, are effective for predicti
ng performance of drastically varying model architectures, and can even generali
ze between model classes. Using these prediction models, we also propose an earl
y stopping method for hyperparameter optimization and meta-modeling, which obtai
ns a speedup of a factor up to 6x in both hyperparameter optimization and meta-m
odeling. Finally, we empirically show that our early stopping method can be seam
lessly incorporated into both reinforcement learning-based architecture selectio
n algorithms and bandit based search methods. Through extensive experimentation,
 we empirically show our performance prediction models and early stopping algori

thm are state-of-the-art in terms of prediction accuracy and speedup achieved while still identifying the optimal model configurations.
****

Latent Space Oddity: on the Curvature of Deep Generative Models
Georgios Arvanitidis,Lars Kai Hansen,Søren Hauberg
Deep generative models provide a systematic way to learn nonlinear data distributions through a set of latent variables and a nonlinear "generator" function that maps latent points into the input space. The nonlinearity of the generator implies that the latent space gives a distorted view of the input space. Under mild conditions, we show that this distortion can be characterized by a stochastic Riemannian metric, and we demonstrate that distances and interpolants are significantly improved under this metric. This in turn improves probability distributions, sampling algorithms and clustering in the latent space. Our geometric analysis further reveals that current generators provide poor variance estimates and we propose a new generator architecture with vastly improved variance estimates. Results are demonstrated on convolutional and fully connected variational autoencoders, but the formalism easily generalizes to other deep generative models.
****

Learning Awareness Models
Brandon Amos,Laurent Dinh,Serkan Cabi,Thomas Rothörl,Sergio Gómez Colmenarejo,Alistair Muldal,Tom Erez,Yuval Tassa,Nando de Freitas,Misha Denil
We consider the setting of an agent with a fixed body interacting with an unknown and uncertain external world. We show that models trained to predict proprioceptive information about the agent's body come to represent objects in the external world. In spite of being trained with only internally available signals, these dynamic body models come to represent external objects through the necessity of predicting their effects on the agent's own body. That is, the model learns holistic persistent representations of objects in the world, even though the only training signals are body signals. Our dynamics model is able to successfully predict distributions over 132 sensor readings over 100 steps into the future and we demonstrate that even when the body is no longer in contact with an object, the latent variables of the dynamics model continue to represent its shape. We show that active data collection by maximizing the entropy of predictions about the body---touch sensors, proprioception and vestibular information---leads to learning of dynamic models that show superior performance when used for control. We also collect data from a real robotic hand and show that the same models can be used to answer questions about properties of objects in the real world. Videos with qualitative results of our models are available at https://goo.gl/mZuqAV.
****

Principled Hybrids of Generative and Discriminative Domain Adaptation
Han Zhao,Zhenyao Zhu,Junjie Hu,Adam Coates,Geoff Gordon
We propose a probabilistic framework for domain adaptation that blends both generative and discriminative modeling in a principled way. Under this framework, generative and discriminative models correspond to specific choices of the prior over parameters. This provides us a very general way to interpolate between generative and discriminative extremes through different choices of priors. By maximizing both the marginal and the conditional log-likelihoods, models derived from this framework can use both labeled instances from the source domain as well as unlabeled instances from \emph{both} source and target domains. Under this framework, we show that the popular reconstruction loss of autoencoder corresponds to an upper bound of the negative marginal log-likelihoods of unlabeled instances, where marginal distributions are given by proper kernel density estimations. This provides a way to interpret the empirical success of autoencoders in domain adaptation and semi-supervised learning. We instantiate our framework using neural networks, and build a concrete model, \emph{DAuto}. Empirically, we demonstrate the effectiveness of DAuto on text, image and speech datasets, showing that it outperforms related competitors when domain adaptation is possible.

****

Towards Binary-Valued Gates for Robust LSTM Training

Zhuohan Li,Di He,Fei Tian,Wei Chen,Tao Qin,Liwei Wang,Tie-Yan Liu
Long Short-Term Memory (LSTM) is one of the most widely used recurrent structures in sequence modeling. Its goal is to use gates to control the information flow (e.g., whether to skip some information/transformation or not) in the recurrent computations, although its practical implementation based on soft gates only partially achieves this goal and is easy to overfit. In this paper, we propose a new way for LSTM training, which pushes the values of the gates towards 0 or 1. By doing so, we can (1) better control the information flow: the gates are mostly open or closed, instead of in a middle state; and (2) avoid overfitting to certain extent: the gates operate at their flat regions, which is shown to correspond to better generalization ability. However, learning towards discrete values of the gates is generally difficult. To tackle this challenge, we leverage the recently developed Gumbel-Softmax trick from the field of variational methods, and make the model trainable with standard backpropagation. Experimental results on language modeling and machine translation show that (1) the values of the gates generated by our method are more reasonable and intuitively interpretable, and (2) our proposed method generalizes better and achieves better accuracy on test sets in all tasks. Moreover, the learnt models are not sensitive to low-precision approximation and low-rank approximation of the gate parameters due to the flat loss surface.
****
GENERATIVE LOW-SHOT NETWORK EXPANSION
Adi Hayat,Mark Kliger,Shachar Fleishman,Daniel Cohen-Or
Conventional deep learning classifiers are static in the sense that they are trained on
a predefined set of classes and learning to classify a novel class typically requires
re-training. In this work, we address the problem of Low-shot network-expansion
learning. We introduce a learning framework which enables expanding a pre-trained
(base) deep network to classify novel classes when the number of examples for the
novel classes is particularly small. We present a simple yet powerful distillation
method where the base network is augmented with additional weights to classify
the novel classes, while keeping the weights of the base network unchanged. We
term this learning hard distillation, since we preserve the response of the network
on the old classes to be equal in both the base and the expanded network. We
show that since only a small number of weights needs to be trained, the hard
distillation excels for low-shot training scenarios. Furthermore, hard distillation
avoids detriment to classification performance on the base classes. Finally, we
show that low-shot network expansion can be done with a very small memory
footprint by using a compact generative model of the base classes training data
with only a negligible degradation relative to learning with the full training set.
****
LEAP: Learning Embeddings for Adaptive Pace
Vithursan Thangarasa,Graham W. Taylor
Determining the optimal order in which data examples are presented to Deep Neural Networks during training is a non-trivial problem. However, choosing a non-trivial scheduling method may drastically improve convergence. In this paper, we propose a Self-Paced Learning (SPL)-fused Deep Metric Learning (DML) framework, which we call Learning Embeddings for Adaptive Pace (LEAP). Our method parameterizes mini-batches dynamically based on the \textit{easiness} and \textit{true diverseness} of the sample within a salient feature representation space. In LEAP, we train an \textit{embedding} Convolutional Neural Network (CNN) to learn an expressive representation space by adaptive density discrimination using the Magnet Loss. The \textit{student} CNN classifier dynamically selects samples to form a

mini-batch based on the \textit{easiness} from cross-entropy losses and \textit{true diverseness} of examples from the representation space sculpted by the \textit{embedding} CNN. We evaluate LEAP using deep CNN architectures for the task of supervised image classification on MNIST, FashionMNIST, CIFAR-10, CIFAR-100, and SVHN. We show that the LEAP framework converges faster with respect to the number of mini-batch updates required to achieve a comparable or better test performance on each of the datasets.
****
The Manifold Assumption and Defenses Against Adversarial Perturbations
Xi Wu,Uyeong Jang,Lingjiao Chen,Somesh Jha
In the adversarial-perturbation problem of neural networks, an adversary starts with a neural network model F and a point $\bfx$ that F classifies correctly, and applies a \emph{small perturbation} to $\bfx$ to produce another point $\bfx'$ that F classifies \emph{incorrectly}. In this paper, we propose taking into account \emph{the inherent confidence information} produced by models when studying adversarial perturbations, where a natural measure of ``confidence'' is \|F(\bfx)\|_\infty$ (i.e. how confident $F$ is about its prediction?). Motivated by a thought experiment based on the manifold assumption, we propose a ``goodness property'' of models which states that \emph{confident regions of a good model should be well separated}. We give formalizations of this property and examine existing robust training objectives in view of them. Interestingly, we find that a recent objective by Madry et al. encourages training a model that satisfies well our formal version of the goodness property, but has a weak control of points that are wrong but with low confidence. However, if Madry et al.'s model is indeed a good solution to their objective, then good and bad points are now distinguishable and we can try to embed uncertain points back to the closest confident region to get (hopefully) correct predictions. We thus propose embedding objectives and algorithms, and perform an empirical study using this method. Our experimental results are encouraging: Madry et al.'s model wrapped with our embedding procedure achieves almost perfect success rate in defending against attacks that the base model fails on, while retaining good generalization behavior.

****
Spectral Graph Wavelets for Structural Role Similarity in Networks
Claire Donnat,Marinka Zitnik,David Hallac,Jure Leskovec
Nodes residing in different parts of a graph can have similar structural roles within their local network topology. The identification of such roles provides key insight into the organization of networks and can also be used to inform machine learning on graphs. However, learning structural representations of nodes is a challenging unsupervised-learning task, which typically involves manually specifying and tailoring topological features for each node. Here we develop GraphWave, a method that represents each node’s local network neighborhood via a low-dimensional embedding by leveraging spectral graph wavelet diffusion patterns. We prove that nodes with similar local network neighborhoods will have similar GraphWave embeddings even though these nodes may reside in very different parts of the network. Our method scales linearly with the number of edges and does not require any hand-tailoring of topological features. We evaluate performance on both synthetic and real-world datasets, obtaining improvements of up to 71% over state-of-the-art baselines.
****
Jointly Learning to Construct and Control Agents using Deep Reinforcement Learning
Charles Schaff,David Yunis,Ayan Chakrabarti,Matthew R. Walter
The physical design of a robot and the policy that controls its motion are inherently coupled. However, existing approaches largely ignore this coupling, instead choosing to alternate between separate design and control phases, which requires expert intuition throughout and risks convergence to suboptimal designs. In this work, we propose a method that jointly optimizes over the physical design of a robot and the corresponding control policy in a model-free fashion, without any need for expert supervision. Given an arbitrary robot morphology, our method

maintains a distribution over the design parameters and uses reinforcement learn ing to train a neural network controller. Throughout training, we refine the rob ot distribution to maximize the expected reward. This results in an assignment t o the robot parameters and neural network policy that are jointly optimal. We ev aluate our approach in the context of legged locomotion, and demonstrate that it discovers novel robot designs and walking gaits for several different morpholog ies, achieving performance comparable to or better than that of hand-crafted des igns.
****

Understanding Local Minima in Neural Networks by Loss Surface Decomposition
Hanock Kwak,Byoung-Tak Zhang
To provide principled ways of designing proper Deep Neural Network (DNN) models, it is essential to understand the loss surface of DNNs under realistic assumpti ons. We introduce interesting aspects for understanding the local minima and ove rall structure of the loss surface. The parameter domain of the loss surface can be decomposed into regions in which activation values (zero or one for rectifie d linear units) are consistent. We found that, in each region, the loss surface have properties similar to that of linear neural networks where every local mini mum is a global minimum. This means that every differentiable local minimum is t he global minimum of the corresponding region. We prove that for a neural networ k with one hidden layer using rectified linear units under realistic assumptions . There are poor regions that lead to poor local minima, and we explain why such regions exist even in the overparameterized DNNs.
****

On Convergence and Stability of GANs
Naveen Kodali,James Hays,Jacob Abernethy,Zsolt Kira
We propose studying GAN training dynamics as regret minimization, which is in co ntrast to the popular view that there is consistent minimization of a divergence between real and generated distributions. We analyze the convergence of GAN tra ining from this new point of view to understand why mode collapse happens. We hy pothesize the existence of undesirable local equilibria in this non-convex game to be responsible for mode collapse. We observe that these local equilibria ofte n exhibit sharp gradients of the discriminator function around some real data po ints. We demonstrate that these degenerate local equilibria can be avoided with a gradient penalty scheme called DRAGAN. We show that DRAGAN enables faster trai ning, achieves improved stability with fewer mode collapses, and leads to genera tor networks with better modeling performance across a variety of architectures and objective functions.
****

Graph Partition Neural Networks for Semi-Supervised Classification
Renjie Liao,Marc Brockschmidt,Daniel Tarlow,Alexander Gaunt,Raquel Urtasun,Richa rd S. Zemel
We present graph partition neural networks (GPNN), an extension of graph neural networks (GNNs) able to handle extremely large graphs. GPNNs alternate between l ocally propagating information between nodes in small subgraphs and globally pro pagating information between the subgraphs. To efficiently partition graphs, we experiment with spectral partitioning and also propose a modified multi-seed flo od fill for fast processing of large scale graphs. We extensively test our model on a variety of semi-supervised node classification tasks. Experimental results indicate that GPNNs are either superior or comparable to state-of-the-art metho ds on a wide variety of datasets for graph-based semi-supervised classification. We also show that GPNNs can achieve similar performance as standard GNNs with f ewer propagation steps.
****

Monotonic Chunkwise Attention
Chung-Cheng Chiu*,Colin Raffel*
Sequence-to-sequence models with soft attention have been successfully applied t o a wide variety of problems, but their decoding process incurs a quadratic time and space cost and is inapplicable to real-time sequence transduction. To addre ss these issues, we propose Monotonic Chunkwise Attention (MoChA), which adaptiv

ely splits the input sequence into small chunks over which soft attention is computed. We show that models utilizing MoChA can be trained efficiently with standard backpropagation while allowing online and linear-time decoding at test time. When applied to online speech recognition, we obtain state-of-the-art results and match the performance of a model using an offline soft attention mechanism. In document summarization experiments where we do not expect monotonic alignments, we show significantly improved performance compared to a baseline monotonic attention-based model.
****

Learning temporal evolution of probability distribution with Recurrent Neural Network
Kyongmin Yeo,Igor Melnyk,Nam Nguyen,Eun Kyung Lee
We propose to tackle a time series regression problem by computing temporal evolution of a probability density function to provide a probabilistic forecast. A Recurrent Neural Network (RNN) based model is employed to learn a nonlinear operator for temporal evolution of a probability density function. We use a softmax layer for a numerical discretization of a smooth probability density functions, which transforms a function approximation problem to a classification task. Explicit and implicit regularization strategies are introduced to impose a smoothness condition on the estimated probability distribution. A Monte Carlo procedure to compute the temporal evolution of the distribution for a multiple-step forecast is presented. The evaluation of the proposed algorithm on three synthetic and two real data sets shows advantage over the compared baselines.
****

Emergent Communication in a Multi-Modal, Multi-Step Referential Game
Katrina Evtimova,Andrew Drozdov,Douwe Kiela,Kyunghyun Cho
Inspired by previous work on emergent communication in referential games, we propose a novel multi-modal, multi-step referential game, where the sender and receiver have access to distinct modalities of an object, and their information exchange is bidirectional and of arbitrary duration. The multi-modal multi-step setting allows agents to develop an internal communication significantly closer to natural language, in that they share a single set of messages, and that the length of the conversation may vary according to the difficulty of the task. We examine these properties empirically using a dataset consisting of images and textual descriptions of mammals, where the agents are tasked with identifying the correct object. Our experiments indicate that a robust and efficient communication protocol emerges, where gradual information exchange informs better predictions and higher communication bandwidth improves generalization.
****

Attacking Binarized Neural Networks
Angus Galloway,Graham W. Taylor,Medhat Moussa
Neural networks with low-precision weights and activations offer compelling
efficiency advantages over their full-precision equivalents. The two most
frequently discussed benefits of quantization are reduced memory consumption,
and a faster forward pass when implemented with efficient bitwise
operations. We propose a third benefit of very low-precision neural networks:
improved robustness against some adversarial attacks, and in the worst case,
performance that is on par with full-precision models. We focus on the very
low-precision case where weights and activations are both quantized to $\pm$1,
and note that stochastically quantizing weights in just one layer can sharply
reduce the impact of iterative attacks. We observe that non-scaled binary neural
networks exhibit a similar effect to the original \emph{defensive distillation}
procedure that led to \emph{gradient masking}, and a false notion of security.
We address this by conducting both black-box and white-box experiments with
binary models that do not artificially mask gradients.
****

DNN Representations as Codewords: Manipulating Statistical Properties via Penalty Regularization
Daeyoung Choi,Changho Shin,Hyunghun Cho,Wonjong Rhee
Performance of Deep Neural Network (DNN) heavily depends on the characteristics

of hidden layer representations. Unlike the codewords of channel coding, however, the representations of learning cannot be directly designed or controlled. Therefore, we develop a family of penalty regularizers where each one aims to affect one of representation's statistical properties such as sparsity, variance, or covariance. The regularizers are extended to perform class-wise regularization, and the extension is found to provide an outstanding shaping capability. A variety of statistical properties are investigated for 10 different regularization strategies including dropout and batch normalization, and several interesting findings are reported. Using the family of regularizers, performance improvements are confirmed for MNIST, CIFAR-100, and CIFAR-10 classification problems. But more importantly, our results suggest that understanding how to manipulate statistical properties of representations can be an important step toward understanding DNN and that the role and effect of DNN regularizers need to be reconsidered.
****

Censoring Representations with Multiple-Adversaries over Random Subspaces
Yusuke Iwasawa,Kotaro Nakayama,Yutaka Matsuo
Adversarial feature learning (AFL) is one of the promising ways for explicitly constrains neural networks to learn desired representations; for example, AFL could help to learn anonymized representations so as to avoid privacy issues. AFL learn such a representations by training the networks to deceive the adversary that predict the sensitive information from the network, and therefore, the success of the AFL heavily relies on the choice of the adversary. This paper proposes a novel design of the adversary, {\em multiple adversaries over random subspaces} (MARS) that instantiate the concept of the {\em volunerableness}. The proposed method is motivated by an assumption that deceiving an adversary could fail to give meaningful information if the adversary is easily fooled, and adversary rely on single classifier suffer from this issues.
In contrast, the proposed method is designed to be less vulnerable, by utilizing the ensemble of independent classifiers where each classifier tries to predict sensitive variables from a different {\em subset} of the representations.
The empirical validations on three user-anonymization tasks show that our proposed method achieves state-of-the-art performances in all three datasets without significantly harming the utility of data.
This is significant because it gives new implications about designing the adversary, which is important to improve the performance of AFL.
****

Auxiliary Guided Autoregressive Variational Autoencoders
Thomas Lucas,Jakob Verbeek
Generative modeling of high-dimensional data is a key problem in machine learning. Successful approaches include latent variable models and autoregressive models. The complementary strengths of these approaches, to model global and local image statistics respectively, suggest hybrid models combining the strengths of both models. Our contribution is to train such hybrid models using an auxiliary loss function that controls which information is captured by the latent variables and what is left to the autoregressive decoder. In contrast, prior work on such hybrid models needed to limit the capacity of the autoregressive decoder to prevent degenerate models that ignore the latent variables and only rely on autoregressive modeling. Our approach results in models with meaningful latent variable representations, and which rely on powerful autoregressive decoders to model image details. Our model generates qualitatively convincing samples, and yields state-of-the-art quantitative results.
****

Sequence Transfer Learning for Neural Decoding
Venkatesh Elango*,Aashish N Patel*,Kai J Miller,Vikash Gilja
A fundamental challenge in designing brain-computer interfaces (BCIs) is decoding behavior from time-varying neural oscillations. In typical applications, decoders are constructed for individual subjects and with limited data leading to restrictions on the types of models that can be utilized. Currently, the best performing decoders are typically linear models capable of utilizing rigid timing constraints with limited training data. Here we demonstrate the use of Long Short-T

erm Memory (LSTM) networks to take advantage of the temporal information present in sequential neural data collected from subjects implanted with electrocortico graphic (ECoG) electrode arrays performing a finger flexion task. Our constructe d models are capable of achieving accuracies that are comparable to existing tec hniques while also being robust to variation in sample data size. Moreover, we u tilize the LSTM networks and an affine transformation layer to construct a novel architecture for transfer learning. We demonstrate that in scenarios where only the affine transform is learned for a new subject, it is possible to achieve re sults comparable to existing state-of-the-art techniques. The notable advantage is the increased stability of the model during training on novel subjects. Relax ing the constraint of only training the affine transformation, we establish our model as capable of exceeding performance of current models across all training data sizes. Overall, this work demonstrates that LSTMs are a versatile model tha t can accurately capture temporal patterns in neural data and can provide a foun dation for transfer learning in neural decoding.
****

The Kanerva Machine: A Generative Distributed Memory
Yan Wu,Greg Wayne,Alex Graves,Timothy Lillicrap
We present an end-to-end trained memory system that quickly adapts to new data a nd generates samples like them. Inspired by Kanerva's sparse distributed memory, it has a robust distributed reading and writing mechanism. The memory is analy tically tractable, which enables optimal on-line compression via a Bayesian upda te-rule. We formulate it as a hierarchical conditional generative model, where m emory provides a rich data-dependent prior distribution. Consequently, the top-d own memory and bottom-up perception are combined to produce the code representin g an observation. Empirically, we demonstrate that the adaptive memory significa ntly improves generative models trained on both the Omniglot and CIFAR datasets. Compared with the Differentiable Neural Computer (DNC) and its variants, our me mory model has greater capacity and is significantly easier to train.
****

Towards Interpretable Chit-chat: Open Domain Dialogue Generation with Dialogue A cts
Wei Wu,Can Xu,Yu Wu,Zhoujun Li
Conventional methods model open domain dialogue generation as a black box throug h end-to-end learning from large scale conversation data. In this work, we make the first step to open the black box by introducing dialogue acts into open doma in dialogue generation. The dialogue acts are generally designed and reveal how people engage in social chat. Inspired by analysis on real data, we propose join tly modeling dialogue act selection and response generation, and perform learnin g with human-human conversations tagged with a dialogue act classifier and a rei nforcement approach to further optimizing the model for long-term conversation. With the dialogue acts, we not only achieve significant improvement over state-o f-the-art methods on response quality for given contexts and long-term conversat ion in both machine-machine simulation and human-machine conversation, but also are capable of explaining why such achievements can be made.
****

Learning how to explain neural networks: PatternNet and PatternAttribution
Pieter-Jan Kindermans,Kristof T. Schütt,Maximilian Alber,Klaus-Robert Müller,Dum itru Erhan,Been Kim,Sven Dähne
DeConvNet, Guided BackProp, LRP, were invented to better understand deep neural networks. We show that these methods do not produce the theoretically correct ex planation for a linear model. Yet they are used on multi-layer networks with mil lions of parameters. This is a cause for concern since linear models are simple neural networks. We argue that explanation methods for neural nets should work r eliably in the limit of simplicity, the linear models. Based on our analysis of linear models we propose a generalization that yields two explanation technique s (PatternNet and PatternAttribution) that are theoretically sound for linear mo dels and produce improved explanations for deep networks.

****

WRPN: Wide Reduced-Precision Networks
Asit Mishra,Eriko Nurvitadhi,Jeffrey J Cook,Debbie Marr
For computer vision applications, prior works have shown the efficacy of reducing numeric precision of model parameters (network weights) in deep neural networks. Activation maps, however, occupy a large memory footprint during both the training and inference step when using mini-batches of inputs. One way to reduce this large memory footprint is to reduce the precision of activations. However, past works have shown that reducing the precision of activations hurts model accuracy. We study schemes to train networks from scratch using reduced-precision activations without hurting accuracy. We reduce the precision of activation maps (along with model parameters) and increase the number of filter maps in a layer, and find that this scheme matches or surpasses the accuracy of the baseline full-precision network. As a result, one can significantly improve the execution efficiency (e.g. reduce dynamic memory footprint, memory band- width and computational energy) and speed up the training and inference process with appropriate hardware support. We call our scheme WRPN -- wide reduced-precision networks. We report results and show that WRPN scheme is better than previously reported accuracies on ILSVRC-12 dataset while being computationally less expensive compared to previously reported reduced-precision networks.
****
Coupled Ensembles of Neural Networks
Anuvabh Dutt,Denis Pellerin,Georges Quénot
We investigate in this paper the architecture of deep convolutional networks. Building on existing state of the art models, we propose a reconfiguration of the model parameters into several parallel branches at the global network level, with each branch being a standalone CNN. We show that this arrangement is an efficient way to significantly reduce the number of parameters while at the same time improving the performance. The use of branches brings an additional form of regularization. In addition to splitting the parameters into parallel branches, we propose a tighter coupling of these branches by averaging their log-probabilities. The tighter coupling favours the learning of better representations, even at the level of the individual branches, as compared to when each branch is trained independently. We refer to this branched architecture as "coupled ensembles". The approach is very generic and can be applied with almost any neural network architecture. With coupled ensembles of DenseNet-BC and parameter budget of 25M, we obtain error rates of 2.92%, 15.68% and 1.50% respectively on CIFAR-10, CIFAR-100 and SVHN tasks. For the same parameter budget, DenseNet-BC has an error rate of 3.46%, 17.18%, and 1.8% respectively. With ensembles of coupled ensembles, of DenseNet-BC networks, with 50M total parameters, we obtain error rates of 2.72%, 15.13% and 1.42% respectively on these tasks.
****
3C-GAN: AN CONDITION-CONTEXT-COMPOSITE GENERATIVE ADVERSARIAL NETWORKS FOR GENERATING IMAGES SEPARATELY
Yeu-Chern Harn,Vladimir Jojic
We present 3C-GAN: a novel multiple generators structures, that contains one conditional generator that generates a semantic part of an image conditional on its input label, and one context generator generates the rest of an image. Compared to original GAN model, this model has multiple generators and gives control over what its generators should generate. Unlike previous multi-generator models use a subsequent generation process, that one layer is generated given the previous layer, our model uses a process of generating different part of the images together. This way the model contains fewer parameters and the generation speed is faster. Speci■cally, the model leverages the label information to separate the object from the image correctly. Since the model conditional on the label information does not restrict to generate other parts of an image, we proposed a cost function that encourages the model to generate only the succinct part of an image in terms of label discrimination. We also found an exclusive prior on the mask of the model help separate the object. The experiments on MNIST, SVHN, and CelebA datasets show 3C-GAN can generate different objects with different generators simultaneously, according to the labels given to each generator.

****
Heterogeneous Bitwidth Binarization in Convolutional Neural Networks
Josh Fromm,Matthai Philipose,Shwetak Patel
Recent work has shown that performing inference with fast, very-low-bitwidth
(e.g., 1 to 2 bits) representations of values in models can yield surprisingly a
ccurate
results. However, although 2-bit approximated networks have been shown to
be quite accurate, 1 bit approximations, which are twice as fast, have restricti
vely
low accuracy. We propose a method to train models whose weights are a mixture
of bitwidths, that allows us to more finely tune the accuracy/speed trade-off. W
e
present the “middle-out” criterion for determining the bitwidth for each value,
and
show how to integrate it into training models with a desired mixture of bitwidth
s.
We evaluate several architectures and binarization techniques on the ImageNet
dataset. We show that our heterogeneous bitwidth approximation achieves superlin
ear
scaling of accuracy with bitwidth. Using an average of only 1.4 bits, we are
able to outperform state-of-the-art 2-bit architectures.
****
Deep Learning is Robust to Massive Label Noise
David Rolnick,Andreas Veit,Serge Belongie,Nir Shavit
Deep neural networks trained on large supervised datasets have led to impressive
 results in recent years. However, since well-annotated datasets can be prohibit
ively expensive and time-consuming to collect, recent work has explored the use
of larger but noisy datasets that can be more easily obtained. In this paper, we
 investigate the behavior of deep neural networks on training sets with massivel
y noisy labels. We show on multiple datasets such as MINST, CIFAR-10 and ImageNe
t that successful learning is possible even with an essentially arbitrary amount
 of noise. For example, on MNIST we find that accuracy of above 90 percent is st
ill attainable even when the dataset has been diluted with 100 noisy examples fo
r each clean example. Such behavior holds across multiple patterns of label nois
e, even when noisy labels are biased towards confusing classes. Further, we show
 how the required dataset size for successful training increases with higher lab
el noise. Finally, we present simple actionable techniques for improving learnin
g in the regime of high label noise.
****
On the difference between building and extracting patterns: a causal analysis of
 deep generative models.
Michel Besserve,Dominik Janzing,Bernhard Schoelkopf
Generative models are important tools to capture and investigate the properties
of complex empirical data. Recent developments such as Generative Adversarial Ne
tworks (GANs) and Variational Auto-Encoders (VAEs) use two very similar, but \te
xtit{reverse}, deep convolutional architectures, one to generate and one to extr
act information from data. Does learning the parameters of both architectures ob
ey the same rules? We exploit the causality principle of independence of mechani
sms to quantify how the weights of successive layers adapt to each other. Using
the recently introduced Spectral Independence Criterion, we quantify the depende
ncies between the kernels of successive convolutional layers and show that those
 are more independent for the generative process than for information extraction
, in line with results from the field of causal inference. In addition, our expe
riments on generation of human faces suggest that more independence between succ
essive layers of generators results in improved performance of these architectur
es.

****
On the Discrimination-Generalization Tradeoff in GANs
Pengchuan Zhang,Qiang Liu,Dengyong Zhou,Tao Xu,Xiaodong He

Generative adversarial training can be generally understood as minimizing certai n moment matching loss defined by a set of discriminator functions, typically n eural networks. The discriminator set should be large enough to be able to uniqu ely identify the true distribution (discriminative), and also be small enough to go beyond memorizing samples (generalizable). In this paper, we show that a dis criminator set is guaranteed to be discriminative whenever its linear span is de nse in the set of bounded continuous functions. This is a very mild condition sa tisfied even by neural networks with a single neuron. Further, we develop genera lization bounds between the learned distribution and true distribution under dif ferent evaluation metrics. When evaluated with neural distance, our bounds show that generalization is guaranteed as long as the discriminator set is small enou gh, regardless of the size of the generator or hypothesis set. When evaluated wi th KL divergence, our bound provides an explanation on the counter-intuitive beh aviors of testing likelihood in GAN training. Our analysis sheds lights on under standing the practical performance of GANs.
****

The Reactor: A fast and sample-efficient Actor-Critic agent for Reinforcement L earning

Audrunas Gruslys,Will Dabney,Mohammad Gheshlaghi Azar,Bilal Piot,Marc Bellemare, Remi Munos

In this work we present a new agent architecture, called Reactor, which combines multiple algorithmic and architectural contributions to produce an agent with h igher sample-efficiency than Prioritized Dueling DQN (Wang et al., 2016) and Cat egorical DQN (Bellemare et al., 2017), while giving better run-time performance than A3C (Mnih et al., 2016). Our first contribution is a new policy evaluation algorithm called Distributional Retrace, which brings multi-step off-policy upda tes to the distributional reinforcement learning setting. The same approach can be used to convert several classes of multi-step policy evaluation algorithms de signed for expected value evaluation into distributional ones. Next, we introduc e the β-leaveone-out policy gradient algorithm which improves the trade-off betw een variance and bias by using action values as a baseline. Our final algorithmi c contribution is a new prioritized replay algorithm for sequences, which exploi ts the temporal locality of neighboring observations for more efficient replay p rioritization. Using the Atari 2600 benchmarks, we show that each of these innov ations contribute to both the sample efficiency and final agent performance. Fin ally, we demonstrate that Reactor reaches state-of-the-art performance after 200 million frames and less than a day of training.
****

Evaluation of generative networks through their data augmentation capacity

Timothée Lesort,Florian Bordes,Jean-Francois Goudou,David Filliat

Generative networks are known to be difficult to assess. Recent works on generat ive models, especially on generative adversarial networks, produce nice samples of varied categories of images. But the validation of their quality is highly de pendent on the method used. A good generator should generate data which contain meaningful and varied information and that fit the distribution of a dataset. Th is paper presents a new method to assess a generator. Our approach is based on t raining a classifier with a mixture of real and generated samples. We train a ge nerative model over a labeled training set, then we use this generative model to sample new data points that we mix with the original training data. This mixtur e of real and generated data is thus used to train a classifier which is afterwa rds tested on a given labeled test dataset. We compare this result with the scor e of the same classifier trained on the real training data mixed with noise. By computing the classifier's accuracy with different ratios of samples from both d istributions (real and generated) we are able to estimate if the generator succe ssfully fits and is able to generalize the distribution of the dataset. Our expe riments compare the result of different generators from the VAE and GAN framewor k on MNIST and fashion MNIST dataset.
****

Towards Synthesizing Complex Programs From Input-Output Examples

Xinyun Chen,Chang Liu,Dawn Song

In recent years, deep learning techniques have been developed to improve the performance of program synthesis from input-output examples. Albeit its significant progress, the programs that can be synthesized by state-of-the-art approaches are still simple in terms of their complexity. In this work, we move a significant step forward along this direction by proposing a new class of challenging tasks in the domain of program synthesis from input-output examples: learning a context-free parser from pairs of input programs and their parse trees. We show that this class of tasks are much more challenging than previously studied tasks, and the test accuracy of existing approaches is almost 0%.

We tackle the challenges by developing three novel techniques inspired by three novel observations, which reveal the key ingredients of using deep learning to synthesize a complex program. First, the use of a non-differentiable machine is the key to effectively restrict the search space. Thus our proposed approach learns a neural program operating a domain-specific non-differentiable machine. Second, recursion is the key to achieve generalizability. Thus, we bake-in the notion of recursion in the design of our non-differentiable machine. Third, reinforcement learning is the key to learn how to operate the non-differentiable machine, but it is also hard to train the model effectively with existing reinforcement learning algorithms from a cold boot. We develop a novel two-phase reinforcement learning-based search algorithm to overcome this issue. In our evaluation, we show that using our novel approach, neural parsing programs can be learned to achieve 100% test accuracy on test inputs that are 500x■ longer than the training samples.
****
Learning non-linear transform with discriminative and minimum information loss priors
Dimche Kostadinov,Slava Voloshynovskiy
This paper proposes a novel approach for learning discriminative and sparse representations. It consists of utilizing two different models. A predefined number of non-linear transform models are used in the learning stage, and one sparsifying transform model is used at test time. The non-linear transform models have discriminative and minimum information loss priors. A novel measure related to the discriminative prior is proposed and defined on the support intersection for the transform representations. The minimum information loss prior is expressed as a constraint on the conditioning and the expected coherence of the transform matrix. An equivalence between the non-linear models and the sparsifying model is shown only when the measure that is used to define the discriminative prior goes to zero. An approximation of the measure used in the discriminative prior is addressed, connecting it to a similarity concentration. To quantify the discriminative properties of the transform representation, we introduce another measure and present its bounds. Reflecting the discriminative quality of the transform representation we name it as discrimination power.

To support and validate the theoretical analysis a practical learning algorithm is presented. We evaluate the advantages and the potential of the proposed algorithm by a computer simulation. A favorable performance is shown considering the execution time, the quality of the representation, measured by the discrimination power and the recognition accuracy in comparison with the state-of-the-art methods of the same category.
****
Image Transformer
Ashish Vaswani,Niki Parmar,Jakob Uszkoreit,Noam Shazeer,Lukasz Kaiser
Image generation has been successfully cast as an autoregressive sequence generation
or transformation problem. Recent work has shown that self-attention is
an effective way of modeling textual sequences. In this work, we generalize a
recently proposed model architecture based on self-attention, the Transformer, to
a sequence modeling formulation of image generation with a tractable likelihood.

By restricting the self-attention mechanism to attend to local neighborhoods we significantly increase the size of images the model can process in practice, despite maintaining significantly larger receptive fields per layer than typical convolutional neural networks. We propose another extension of self-attention allowing it to efficiently take advantage of the two-dimensional nature of images. While conceptually simple, our generative models trained on two image data sets are competitive with or significantly outperform the current state of the art in autoregressive image generation on two different data sets, CIFAR-10 and ImageNet. We also present results on image super-resolution with a large magnification ratio, applying an encoder-decoder configuration of our architecture. In a human evaluation study, we show that our super-resolution models improve significantly over previously published autoregressive super-resolution models. Images they generate fool human observers three times more often than the previous state of the art.

****

Do GANs learn the distribution? Some Theory and Empirics

Sanjeev Arora,Andrej Risteski,Yi Zhang

Do GANS (Generative Adversarial Nets) actually learn the target distribution? The foundational paper of Goodfellow et al. (2014) suggested they do, if they were given sufficiently large deep nets, sample size, and computation time. A recent theoretical analysis in Arora et al. (2017) raised doubts whether the same holds when discriminator has bounded size. It showed that the training objective can approach its optimum value even if the generated distribution has very low support. In other words, the training objective is unable to prevent mode collapse. The current paper makes two contributions. (1) It proposes a novel test for estimating support size using the birthday paradox of discrete probability. Using this evidence is presented that well-known GANs approaches do learn distributions of fairly low support. (2) It theoretically studies encoder-decoder GANs architectures (e.g., BiGAN/ALI), which were proposed to learn more meaningful features via GANs, and consequently to also solve the mode-collapse issue. Our result shows that such encoder-decoder training objectives also cannot guarantee learning of the full distribution because they cannot prevent serious mode collapse. More seriously, they cannot prevent learning meaningless codes for data, contrary to usual intuition.

****

Sparse Attentive Backtracking: Long-Range Credit Assignment in Recurrent Networks

Nan Rosemary Ke,Anirudh Goyal,Olexa Bilaniuk,Jonathan Binas,Laurent Charlin,Chris Pal,Yoshua Bengio

A major drawback of backpropagation through time (BPTT) is the difficulty of learning long-term dependencies, coming from having to propagate credit information backwards through every single step of the forward computation. This makes BPTT both computationally impractical and biologically implausible. For this reason, full backpropagation through time is rarely used on long sequences, and truncated backpropagation through time is used as a heuristic. However, this usually leads to biased estimates of the gradient in which longer term dependencies are ignored. Addressing this issue, we propose an alternative algorithm, Sparse Attentive Backtracking, which might also be related to principles used by brains to learn long-term dependencies. Sparse Attentive Backtracking learns an attention mechanism over the hidden states of the past and selectively backpropagates through paths with high attention weights. This allows the model to learn long term dependencies while only backtracking for a small number of time steps, not just from the recent past but also from attended relevant past states.

****

Learning One-hidden-layer Neural Networks with Landscape Design

Rong Ge,Jason D. Lee,Tengyu Ma

We consider the problem of learning a one-hidden-layer neural network: we assume the input x is from Gaussian distribution and the label $y = a \sigma(Bx) + \xi$, where a is a nonnegative vector and B is a full-rank weight matrix, and ξ is a noise vector. We first give an analytic formula for the population risk of the standard squared loss and demonstrate that it implicitly attempts to decompose a sequence of low-rank tensors simultaneously.
■
Inspired by the formula, we design a non-convex objective function G whose landscape is guaranteed to have the following properties:■

1. All local minima of G are also global minima.
2. All global minima of G correspond to the ground truth parameters.
3. The value and gradient of G can be estimated using samples.

■
With these properties, stochastic gradient descent on G provably converges to the global minimum and learn the ground-truth parameters. We also prove finite sample complexity results and validate the results by simulations.
****
Shifting Mean Activation Towards Zero with Bipolar Activation Functions
Lars Hiller Eidnes,Arild Nøkland
We propose a simple extension to the ReLU-family of activation functions that allows them to shift the mean activation across a layer towards zero. Combined with proper weight initialization, this alleviates the need for normalization layers. We explore the training of deep vanilla recurrent neural networks (RNNs) with up to 144 layers, and show that bipolar activation functions help learning in this setting. On the Penn Treebank and Text8 language modeling tasks we obtain competitive results, improving on the best reported results for non-gated networks. In experiments with convolutional neural networks without batch normalization, we find that bipolar activations produce a faster drop in training error, and results in a lower test error on the CIFAR-10 classification task.
****
Baseline-corrected space-by-time non-negative matrix factorization for decoding single trial population spike trains
Arezoo Alizadeh,Marion Mutter,Thomas Münch,Arno Onken,Stefano Panzeri
Activity of populations of sensory neurons carries stimulus information in both the temporal and the spatial dimensions. This poses the question of how to compactly represent all the information that the population codes carry across all these dimensions. Here, we developed an analytical method to factorize a large number of retinal ganglion cells' spike trains into a robust low-dimensional representation that captures efficiently both their spatial and temporal information. In particular, we extended previously used single-trial space-by-time tensor decomposition based on non-negative matrix factorization to efficiently discount pre-stimulus baseline activity. On data recorded from retinal ganglion cells with strong pre-stimulus baseline, we showed that in situations were the stimulus elicits a strong change in firing rate, our extensions yield a boost in stimulus decoding performance. Our results thus suggest that taking into account the baseline can be important for finding a compact information-rich representation of neural activity.
****
Parametric Adversarial Divergences are Good Task Losses for Generative Modeling
Gabriel Huang,Hugo Berard,Ahmed Touati,Gauthier Gidel,Pascal Vincent,Simon Lacoste-Julien
Generative modeling of high dimensional data like images is a notoriously difficult and ill-defined problem. In particular, how to evaluate a learned generative model is unclear.
In this paper, we argue that *adversarial learning*, pioneered with generative adversarial networks (GANs), provides an interesting framework to implicitly define more meaningful task losses for unsupervised tasks, such as for generating "visually realistic" images. By relating GANs and structured prediction under the framework of statistical decision theory, we put into light links between recent

advances in structured prediction theory and the choice of the divergence in GANs. We argue that the insights about the notions of "hard" and "easy" to learn losses can be analogously extended to adversarial divergences. We also discuss the attractive properties of parametric adversarial divergences for generative modeling, and perform experiments to show the importance of choosing a divergence that reflects the final task.
****
Parametrizing filters of a CNN with a GAN
Yannic Kilcher,Gary Becigneul,Thomas Hofmann
It is commonly agreed that the use of relevant invariances as a good statistical bias is important in machine-learning. However, most approaches that explicitely incorporate invariances into a model architecture only make use of very simple transformations, such as translations and rotations. Hence, there is a need for methods to model and extract richer transformations that capture much higher-level invariances. To that end, we introduce a tool allowing to parametrize the set of filters of a trained convolutional neural network with the latent space of a generative adversarial network. We then show that the method can capture highly non-linear invariances of the data by visualizing their effect in the data space.
****
Dissecting Adam: The Sign, Magnitude and Variance of Stochastic Gradients
Lukas Balles,Philipp Hennig
The ADAM optimizer is exceedingly popular in the deep learning community. Often it works very well, sometimes it doesn’t. Why? We interpret ADAM as a combination of two aspects: for each weight, the update direction is determined by the sign of the stochastic gradient, whereas the update magnitude is solely determined by an estimate of its relative variance. We disentangle these two aspects and analyze them in isolation, shedding light on ADAM ’s inner workings. Transferring the "variance adaptation” to momentum- SGD gives rise to a novel method, completing the practitioner’s toolbox for problems where ADAM fails.
****
Unbiased scalable softmax optimization
Francois Fagan,Garud Iyengar
Recent neural network and language models have begun to rely on softmax distributions with an extremely large number of categories. In this context calculating the softmax normalizing constant is prohibitively expensive. This has spurred a growing literature of efficiently computable but biased estimates of the softmax. In this paper we present the first two unbiased algorithms for maximizing the softmax likelihood whose work per iteration is independent of the number of classes and datapoints (and does not require extra work at the end of each epoch). We compare our unbiased methods' empirical performance to the state-of-the-art on seven real world datasets, where they comprehensively outperform all competitors.
****
Toward predictive machine learning for active vision
Emmanuel Daucé
We develop a comprehensive description of the active inference framework, as proposed by Friston (2010), under a machine-learning compliant perspective. Stemming from a biological inspiration and the auto-encoding principles, a sketch of a cognitive architecture is proposed that should provide ways to implement estimation-oriented control policies. Computer simulations illustrate the effectiveness of the approach through a foveated inspection of the input data. The pros and cons of the control policy are analyzed in detail, showing interesting promises in terms of processing compression. Though optimizing future posterior entropy over the actions set is shown enough to attain locally optimal action selection, offline calculation using class-specific saliency maps is shown better for it saves processing costs through saccades pathways pre-processing, with a negligible effect on the recognition/compression rates.
****
PDE-Net: Learning PDEs from Data

Zichao Long,Yiping Lu,Xianzhong Ma,Bin Dong
Partial differential equations (PDEs) play a prominent role in many disciplines
 such as applied mathematics, physics, chemistry, material science, computer sci
ence, etc. PDEs are commonly derived based on physical laws or empirical observa
tions. However, the governing equations for many complex systems in modern appli
cations are still not fully known. With the rapid development of sensors, comput
ational power, and data storage in the past decade, huge quantities of data can
be easily collected and efficiently stored. Such vast quantity of data offers ne
w opportunities for data-driven discovery of hidden physical laws. Inspired by t
he latest development of neural network designs in deep learning, we propose a n
ew feed-forward deep network, called PDE-Net, to fulfill two objectives at the s
ame time: to accurately predict dynamics of complex systems and to uncover the u
nderlying hidden PDE models. The basic idea of the proposed PDE-Net is to learn
differential operators by learning convolution kernels (filters), and apply neur
al networks or other machine learning methods to approximate the unknown nonline
ar responses. Comparing with existing approaches, which either assume the form o
f the nonlinear response is known or fix certain finite difference approximation
s of differential operators, our approach has the most flexibility by learning b
oth differential operators and the nonlinear responses. A special feature of the
 proposed PDE-Net is that all filters are properly constrained, which enables us
 to easily identify the governing PDE models while still maintaining the express
ive and predictive power of the network. These constrains are carefully designed
 by fully exploiting the relation between the orders of differential operators a
nd the orders of sum rules of filters (an important concept originated from wave
let theory). We also discuss relations of the PDE-Net with some existing network
s in computer vision such as Network-In-Network (NIN) and Residual Neural Networ
k (ResNet). Numerical experiments show that the PDE-Net has the potential to unc
over the hidden PDE of the observed dynamics, and predict the dynamical behavior
 for a relatively long time, even in a noisy environment.
****
Image Segmentation by Iterative Inference from Conditional Score Estimation
Adriana Romero,Michal Drozdzal,Akram Erraqabi,Simon Jégou,Yoshua Bengio
Inspired by the combination of feedforward and iterative computations in the vis
ual cortex, and taking advantage of the ability of denoising autoencoders to est
imate the score of a joint distribution, we propose a novel approach to iterativ
e inference for capturing and exploiting the complex joint distribution of outpu
t variables conditioned on some input variables. This approach is applied to ima
ge pixel-wise segmentation, with the estimated conditional score used to perform
 gradient ascent towards a mode of the estimated conditional distribution. This
extends previous work on score estimation by denoising autoencoders to the case
of a conditional distribution, with a novel use of a corrupted feedforward predi
ctor replacing Gaussian corruption. An advantage of this approach over more clas
sical ways to perform iterative inference for structured outputs, like condition
al random fields (CRFs), is that it is not any more necessary to define an expli
cit energy function linking the output variables. To keep computations tractable
, such energy function parametrizations are typically fairly constrained, involv
ing only a few neighbors of each of the output variables in each clique. We expe
rimentally find that the proposed iterative inference from conditional score est
imation by conditional denoising autoencoders performs better than comparable mo
dels based on CRFs or those not using any explicit modeling of the conditional j
oint distribution of outputs.
****
Parametric Information Bottleneck to Optimize Stochastic Neural Networks
Thanh T. Nguyen,Jaesik Choi
In this paper, we present a layer-wise learning of stochastic neural networks (S
NNs) in an information-theoretic perspective. In each layer of an SNN, the compr
ession and the relevance are defined to quantify the amount of information that
the layer contains about the input space and the target space, respectively. We
jointly optimize the compression and the relevance of all parameters in an SNN t
o better exploit the neural network's representation. Previously, the Informatio

n Bottleneck (IB) framework (\cite{Tishby99}) extracts relevant information for a target variable. Here, we propose Parametric Information Bottleneck (PIB) for a neural network by utilizing (only) its model parameters explicitly to approximate the compression and the relevance. We show that, as compared to the maximum likelihood estimate (MLE) principle, PIBs : (i) improve the generalization of neural networks in classification tasks, (ii) push the representation of neural networks closer to the optimal information-theoretical representation in a faster manner.
****

DEEP DENSITY NETWORKS AND UNCERTAINTY IN RECOMMENDER SYSTEMS

Yoel Zeldes,Stavros Theodorakis,Efrat Solodnik,Aviv Rotman,Gil Chamiel,Dan Friedman

Building robust online content recommendation systems requires learning com- plex interactions between user preferences and content features. The field has evolved rapidly in recent years from traditional multi-arm bandit and collabora- tive filtering techniques, with new methods integrating Deep Learning models that enable to capture non-linear feature interactions. Despite progress, the dynamic nature of online recommendations still poses great challenges, such as finding the delicate balance between exploration and exploitation. In this paper we provide a novel method, Deep Density Networks (DDN) which deconvolves measurement and data uncertainty and predicts probability densities of CTR, enabling us to perform more efficient exploration of the feature space. We show the usefulness of using DDN online in a real world content recommendation system that serves billions of recommendations per day, and present online and offline results to eval- uate the benefit of using DDN.
****

The Variational Homoencoder: Learning to Infer High-Capacity Generative Models from Few Examples

Luke Hewitt,Andrea Gane,Tommi Jaakkola,Joshua B. Tenenbaum

Hierarchical Bayesian methods have the potential to unify many related tasks (e.g. k-shot classification, conditional, and unconditional generation) by framing each as inference within a single generative model. We show that existing approaches for learning such models can fail on expressive generative networks such as PixelCNNs, by describing the global distribution with little reliance on latent variables. To address this, we develop a modification of the Variational Autoencoder in which encoded observations are decoded to new elements from the same class; the result, which we call a Variational Homoencoder (VHE), may be understood as training a hierarchical latent variable model which better utilises latent variables in these cases. Using this framework enables us to train a hierarchical PixelCNN for the Omniglot dataset, outperforming all existing models on test set likelihood. With a single model we achieve both strong one-shot generation and near human-level classification, competitive with state-of-the-art discriminative classifiers. The VHE objective extends naturally to richer dataset structures such as factorial or hierarchical categories, as we illustrate by training models to separate character content from simple variations in drawing style, and to generalise the style of an alphabet to new characters.
****

Cross-Corpus Training with TreeLSTM for the Extraction of Biomedical Relationships from Text

Legrand Joël,Yannick Toussaint,Chedy Raïssi,Adrien Coulet

A bottleneck problem in machine learning-based relationship extraction (RE) algorithms, and particularly of deep learning-based ones, is the availability of training data in the form of annotated corpora. For specific domains, such as biomedicine, the long time and high expertise required for the development of manually annotated corpora explain that most of the existing one are relatively small (i.e., hundreds of sentences). Beside, larger corpora focusing on general or domain-specific relationships (such as citizenship or drug-drug interactions) have been developed. In this paper, we study how large annotated corpora developed for alternative tasks may improve the performances on biomedicine related tasks, for which few annotated resources are available. We experiment two deep learning-b

ased models to extract relationships from biomedical texts with high performance. The first one combine locally extracted features using a Convolutional Neural Network (CNN) model, while the second exploit the syntactic structure of sentences using a Recursive Neural Network (RNN) architecture. Our experiments show that, contrary to the former, the latter benefits from a cross-corpus learning strategy to improve the performance of relationship extraction tasks. Indeed our approach leads to the best published performances for two biomedical RE tasks, and to state-of-the-art results for two other biomedical RE tasks, for which few annotated resources are available (less than 400 manually annotated sentences). This may be particularly impactful in specialized domains in which training resources are scarce, because they would benefit from the training data of other domains for which large annotated corpora does exist.
****

DNN Feature Map Compression using Learned Representation over GF(2)

Denis A. Gudovskiy,Alec Hodgkinson,Luca Rigazio

In this paper, we introduce a method to compress intermediate feature maps of deep neural networks (DNNs) to decrease memory storage and bandwidth requirements during inference. Unlike previous works, the proposed method is based on converting fixed-point activations into vectors over the smallest GF(2) finite field followed by nonlinear dimensionality reduction (NDR) layers embedded into a DNN. Such an end-to-end learned representation finds more compact feature maps by exploiting quantization redundancies within the fixed-point activations along the channel or spatial dimensions. We apply the proposed network architecture to the tasks of ImageNet classification and PASCAL VOC object detection. Compared to prior approaches, the conducted experiments show a factor of 2 decrease in memory requirements with minor degradation in accuracy while adding only bitwise computations.
****

Learning From Noisy Singly-labeled Data

Ashish Khetan,Zachary C. Lipton,Animashree Anandkumar

Supervised learning depends on annotated examples, which are taken to be the ground truth. But these labels often come from noisy crowdsourcing platforms, like Amazon Mechanical Turk. Practitioners typically collect multiple labels per example and aggregate the results to mitigate noise (the classic crowdsourcing problem). Given a fixed annotation budget and unlimited unlabeled data, redundant annotation comes at the expense of fewer labeled examples. This raises two fundamental questions: (1) How can we best learn from noisy workers? (2) How should we allocate our labeling budget to maximize the performance of a classifier? We propose a new algorithm for jointly modeling labels and worker quality from noisy crowd-sourced data. The alternating minimization proceeds in rounds, estimating worker quality from disagreement with the current model and then updating the model by optimizing a loss function that accounts for the current estimate of worker quality. Unlike previous approaches, even with only one annotation per example, our algorithm can estimate worker quality. We establish a generalization error bound for models learned with our algorithm and establish theoretically that it's better to label many examples once (vs less multiply) when worker quality exceeds a threshold. Experiments conducted on both ImageNet (with simulated noisy workers) and MS-COCO (using the real crowdsourced labels) confirm our algorithm's benefits.
****

Understanding Short-Horizon Bias in Stochastic Meta-Optimization

Yuhuai Wu,Mengye Ren,Renjie Liao,Roger Grosse.

Careful tuning of the learning rate, or even schedules thereof, can be crucial to effective neural net training. There has been much recent interest in gradient-based meta-optimization, where one tunes hyperparameters, or even learns an optimizer, in order to minimize the expected loss when the training procedure is unrolled. But because the training procedure must be unrolled thousands of times, the meta-objective must be defined with an orders-of-magnitude shorter time horizon than is typical for neural net training. We show that such short-horizon meta-objectives cause a serious bias towards small step sizes, an effect we term sh

ort-horizon bias. We introduce a toy problem, a noisy quadratic cost function, on which we analyze short-horizon bias by deriving and comparing the optimal schedules for short and long time horizons. We then run meta-optimization experiments (both offline and online) on standard benchmark datasets, showing that meta-optimization chooses too small a learning rate by multiple orders of magnitude, even when run with a moderately long time horizon (100 steps) typical of work in the area. We believe short-horizon bias is a fundamental problem that needs to be addressed if meta-optimization is to scale to practical neural net training regimes.
****
Ensemble Robustness and Generalization of Stochastic Deep Learning Algorithms
Tom Zahavy,Bingyi Kang,Alex Sivak,Jiashi Feng,Huan Xu,Shie Mannor
The question why deep learning algorithms generalize so well has attracted increasing
research interest. However, most of the well-established approaches,
such as hypothesis capacity, stability or sparseness, have not provided complete
explanations (Zhang et al., 2016; Kawaguchi et al., 2017). In this work, we focus
on the robustness approach (Xu & Mannor, 2012), i.e., if the error of a hypothesis
will not change much due to perturbations of its training examples, then it
will also generalize well. As most deep learning algorithms are stochastic (e.g.,
Stochastic Gradient Descent, Dropout, and Bayes-by-backprop), we revisit the robustness
arguments of Xu & Mannor, and introduce a new approach – ensemble
robustness – that concerns the robustness of a population of hypotheses. Through
the lens of ensemble robustness, we reveal that a stochastic learning algorithm can
generalize well as long as its sensitiveness to adversarial perturbations is bounded
in average over training examples. Moreover, an algorithm may be sensitive to
some adversarial examples (Goodfellow et al., 2015) but still generalize well. To
support our claims, we provide extensive simulations for different deep learning
algorithms and different network architectures exhibiting a strong correlation between
ensemble robustness and the ability to generalize.
****
A Scalable Laplace Approximation for Neural Networks
Hippolyt Ritter,Aleksandar Botev,David Barber
We leverage recent insights from second-order optimisation for neural networks to construct a Kronecker factored Laplace approximation to the posterior over the weights of a trained network. Our approximation requires no modification of the training procedure, enabling practitioners to estimate the uncertainty of their models currently used in production without having to retrain them. We extensively compare our method to using Dropout and a diagonal Laplace approximation for estimating the uncertainty of a network. We demonstrate that our Kronecker factored method leads to better uncertainty estimates on out-of-distribution data and is more robust to simple adversarial attacks. Our approach only requires calculating two square curvature factor matrices for each layer. Their size is equal to the respective square of the input and output size of the layer, making the method efficient both computationally and in terms of memory usage. We illustrate its scalability by applying it to a state-of-the-art convolutional network architecture.
****
Learning Deep Generative Models With Discrete Latent Variables
Hengyuan Hu,Ruslan Salakhutdinov
There have been numerous recent advancements on learning deep generative models with latent variables thanks to the reparameterization trick that allows to trai

n deep directed models effectively. However, since reparameterization trick onl
y works on continuous variables, deep generative models with discrete latent var
iables still remain hard to train and perform considerably worse than their cont
inuous counterparts. In this paper, we attempt to shrink this gap by introducing
 a new architecture and its learning procedure. We develop a hybrid generative
model with binary latent variables that consists of an undirected graphical mode
l and a deep neural network. We propose an efficient two-stage pretraining and t
raining procedure that is crucial for learning these models. Experiments on bina
rized digits and images of natural scenes demonstrate that our model achieves cl
ose to the state-of-the-art performance in terms of density estimation and is ca
pable of generating coherent images of natural scenes.
****
A PAC-Bayesian Approach to Spectrally-Normalized Margin Bounds for Neural Networ
ks
Behnam Neyshabur,Srinadh Bhojanapalli,Nathan Srebro
We present a generalization bound for feedforward neural networks in terms of th
e product of the spectral norm of the layers and the Frobenius norm of the weigh
ts. The generalization bound is derived using a PAC-Bayes analysis.
****
Deep Learning as a Mixed Convex-Combinatorial Optimization Problem
Abram L. Friesen,Pedro Domingos
As neural networks grow deeper and wider, learning networks with hard-threshold
activations is becoming increasingly important, both for network quantization, w
hich can drastically reduce time and energy requirements, and for creating large
 integrated systems of deep networks, which may have non-differentiable componen
ts and must avoid vanishing and exploding gradients for effective learning. Howe
ver, since gradient descent is not applicable to hard-threshold functions, it is
 not clear how to learn them in a principled way. We address this problem by obs
erving that setting targets for hard-threshold hidden units in order to minimize
 loss is a discrete optimization problem, and can be solved as such. The discret
e optimization goal is to find a set of targets such that each unit, including t
he output, has a linearly separable problem to solve. Given these targets, the n
etwork decomposes into individual perceptrons, which can then be learned with st
andard convex approaches. Based on this, we develop a recursive mini-batch algor
ithm for learning deep hard-threshold networks that includes the popular but poo
rly justified straight-through estimator as a special case. Empirically, we show
 that our algorithm improves classification accuracy in a number of settings, in
cluding for AlexNet and ResNet-18 on ImageNet, when compared to the straight-thr
ough estimator.
****
Recurrent Neural Networks with Top-k Gains for Session-based Recommendations
Balázs Hidasi,Alexandros Karatzoglou
RNNs have been shown to be excellent models for sequential data and in particula
r for session-based user behavior. The use of RNNs provides impressive performan
ce benefits over classical methods in session-based recommendations. In this wor
k we introduce a novel ranking loss function tailored for RNNs in recommendation
 settings. The better performance of such loss over alternatives, along with fur
ther tricks and improvements described in this work, allow to achieve an overall
 improvement of up to 35% in terms of MRR and Recall@20 over previous session-ba
sed RNN solutions and up to 51% over classical collaborative filtering approache
s. Unlike data augmentation-based improvements, our method does not increase tra
ining times significantly.
****
Graph Classification with 2D Convolutional Neural Networks
Antoine J.-P. Tixier,Giannis Nikolentzos,Polykarpos Meladianos,Michalis Vazirgia
nnis
Graph classification is currently dominated by graph kernels, which, while power
ful, suffer some significant limitations. Convolutional Neural Networks (CNNs) o
ffer a very appealing alternative. However, processing graphs with CNNs is not t
rivial. To address this challenge, many sophisticated extensions of CNNs have re

cently been proposed. In this paper, we reverse the problem: rather than proposing yet another graph CNN model, we introduce a novel way to represent graphs as multi-channel image-like structures that allows them to be handled by vanilla 2D CNNs. Despite its simplicity, our method proves very competitive to state-of-the-art graph kernels and graph CNNs, and outperforms them by a wide margin on some datasets. It is also preferable to graph kernels in terms of time complexity. Code and data are publicly available.
****
Machine Learning by Two-Dimensional Hierarchical Tensor Networks: A Quantum Information Theoretic Perspective on Deep Architectures
Ding Liu,Shi-Ju Ran,Peter Wittek,Cheng Peng,Raul Blázquez García,Gang Su,Maciej Lewenstein
The resemblance between the methods used in studying quantum-many body physics and in machine learning has drawn considerable attention. In particular, tensor networks (TNs) and deep learning architectures bear striking similarities to the extent that TNs can be used for machine learning. Previous results used one-dimensional TNs in image recognition, showing limited scalability and a request of high bond dimension. In this work, we train two-dimensional hierarchical TNs to solve image recognition problems, using a training algorithm derived from the multipartite entanglement renormalization ansatz (MERA). This approach overcomes scalability issues and implies novel mathematical connections among quantum many-body physics, quantum information theory, and machine learning. While keeping the TN unitary in the training phase, TN states can be defined, which optimally encodes each class of the images into a quantum many-body state. We study the quantum features of the TN states, including quantum entanglement and fidelity. We suggest these quantities could be novel properties that characterize the image classes, as well as the machine learning tasks. Our work could be further applied to identifying possible quantum properties of certain artificial intelligence methods.
****
Learning to play slot cars and Atari 2600 games in just minutes
Lionel Cordesses,Omar Bentahar,Julien Page
Machine learning algorithms for controlling devices will need to learn quickly, with few trials. Such a goal can be attained with concepts borrowed from continental philosophy and formalized using tools from the mathematical theory of categories. Illustrations of this approach are presented on a cyberphysical system: the slot car game, and also on Atari 2600 games.
****
Optimizing the Latent Space of Generative Networks
Piotr Bojanowski,Armand Joulin,David Lopez-Paz,Arthur Szlam
Generative Adversarial Networks (GANs) have achieved remarkable results in the task of generating realistic natural images. In most applications, GAN models share two aspects in common. On the one hand, GANs training involves solving a challenging saddle point optimization problem, interpreted as an adversarial game between a generator and a discriminator functions. On the other hand, the generator and the discriminator are parametrized in terms of deep convolutional neural networks. The goal of this paper is to disentangle the contribution of these two factors to the success of GANs. In particular, we introduce Generative Latent Optimization (GLO), a framework to train deep convolutional generators without using discriminators, thus avoiding the instability of adversarial optimization problems. Throughout a variety of experiments, we show that GLO enjoys many of the desirable properties of GANs: learning from large data, synthesizing visually-appealing samples, interpolating meaningfully between samples, and performing linear arithmetic with noise vectors.
****
Simple Fast Convolutional Feature Learning
David Macêdo,Cleber Zanchettin,Teresa Ludermir
The quality of the features used in visual recognition is of fundamental importance for the overall system. For a long time, low-level hand-designed feature algorithms as SIFT and HOG have obtained the best results on image recognition. Vis

ual features have recently been extracted from trained convolutional neural netw
orks. Despite the high-quality results, one of the main drawbacks of this approa
ch, when compared with hand-designed features, is the training time required dur
ing the learning process. In this paper, we propose a simple and fast way to tra
in supervised convolutional models to feature extraction while still maintaining
 its high-quality. This methodology is evaluated on different datasets and compa
red with state-of-the-art approaches.
****
Leveraging Grammar and Reinforcement Learning for Neural Program Synthesis
Rudy Bunel,Matthew Hausknecht,Jacob Devlin,Rishabh Singh,Pushmeet Kohli
Program synthesis is the task of automatically generating a program consistent w
ith
a specification. Recent years have seen proposal of a number of neural approache
s
for program synthesis, many of which adopt a sequence generation paradigm simila
r
to neural machine translation, in which sequence-to-sequence models are trained
to
maximize the likelihood of known reference programs. While achieving impressive
results, this strategy has two key limitations. First, it ignores Program Aliasi
ng: the
fact that many different programs may satisfy a given specification (especially
with
incomplete specifications such as a few input-output examples). By maximizing
the likelihood of only a single reference program, it penalizes many semanticall
y
correct programs, which can adversely affect the synthesizer performance. Second
,
this strategy overlooks the fact that programs have a strict syntax that can be
efficiently checked. To address the first limitation, we perform reinforcement
learning on top of a supervised model with an objective that explicitly maximize
s
the likelihood of generating semantically correct programs. For addressing the
second limitation, we introduce a training procedure that directly maximizes the
probability of generating syntactically correct programs that fulfill the specif
ication.
We show that our contributions lead to improved accuracy of the models, especial
ly
in cases where the training data is limited.
****
Exponentially vanishing sub-optimal local minima in multilayer neural networks
Daniel Soudry,Elad Hoffer
Background: Statistical mechanics results (Dauphin et al. (2014); Choromanska et
 al. (2015)) suggest that local minima with high error are exponentially rare in
 high dimensions. However, to prove low error guarantees for Multilayer Neural N
etworks (MNNs), previous works so far required either a heavily modified MNN mod
el or training method, strong assumptions on the labels (e.g., “near” linear sep
arability), or an unrealistically wide hidden layer with \Omega\(N) units.

Results: We examine a MNN with one hidden layer of piecewise linear units, a sin
gle output, and a quadratic loss. We prove that, with high probability in the li
mit of N\rightarrow\infty datapoints, the volume of differentiable regions of th
e empiric loss containing sub-optimal differentiable local minima is exponential
ly vanishing in comparison with the same volume of global minima, given standard
 normal input of dimension d_0=\tilde{\Omega}(\sqrt{N}), and a more realistic nu
mber of d_1=\tilde{\Omega}(N/d_0) hidden units. We demonstrate our results numer
ically: for example, 0% binary classification training error on CIFAR with only
N/d_0 = 16 hidden neurons.
****
LEARNING TO ORGANIZE KNOWLEDGE WITH N-GRAM MACHINES

Fan Yang,Jiazhong Nie,William W. Cohen,Ni Lao
Deep neural networks (DNNs) had great success on NLP tasks such as language modeling, machine translation and certain question answering (QA) tasks. However, the success is limited at more knowledge intensive tasks such as QA from a big corpus. Existing end-to-end deep QA models (Miller et al., 2016; Weston et al., 2014) need to read the entire text after observing the question, and therefore their complexity in responding a question is linear in the text size. This is prohibitive for practical tasks such as QA from Wikipedia, a novel, or the Web. We propose to solve this scalability issue by using symbolic meaning representations, which can be indexed and retrieved efficiently with complexity that is independent of the text size. More specifically, we use sequence-to-sequence models to encode knowledge symbolically and generate programs to answer questions from the encoded knowledge. We apply our approach, called the N-Gram Machine (NGM), to the bAbI tasks (Weston et al., 2015) and a special version of them (“life-long bAbI”) which has stories of up to 10 million sentences. Our experiments show that NGM can successfully solve both of these tasks accurately and efficiently. Unlike fully differentiable memory models, NGM’s time complexity and answering quality are not affected by the story length. The whole system of NGM is trained end-to-end with REINFORCE (Williams, 1992). To avoid high variance in gradient estimation, which is typical in discrete latent variable models, we use beam search instead of sampling. To tackle the exponentially large search space, we use a stabilized auto-encoding objective and a structure tweak procedure to iteratively reduce and refine the search space.

****

Automatic Goal Generation for Reinforcement Learning Agents
David Held,Xinyang Geng,Carlos Florensa,Pieter Abbeel
Reinforcement learning (RL) is a powerful technique to train an agent to perform a task. However, an agent that is trained using RL is only capable of achieving the single task that is specified via its reward function.   Such an approach does not scale well to settings in which an agent needs to perform a diverse set of tasks, such as navigating to varying positions in a room or moving objects to varying locations. Instead, we propose a method that allows an agent to automatically discover the range of tasks that it is capable of performing in its environment. We use a generator network to propose tasks for the agent to try to achieve, each task being specified as reaching a certain parametrized subset of the state-space. The generator network is optimized using adversarial training to produce tasks that are always at the appropriate level of difficulty for the agent. Our method thus automatically produces a curriculum of tasks for the agent to learn. We show that, by using this framework, an agent can efficiently and automatically learn to perform a wide set of tasks without requiring any prior knowledge of its environment (Videos and code available at: https://sites.google.com/view/goalgeneration4rl). Our method can also learn to achieve tasks with sparse rewards, which pose significant challenges for traditional RL methods.
****
Unbiased Online Recurrent Optimization
Corentin Tallec,Yann Ollivier
The novel \emph{Unbiased Online Recurrent Optimization} (UORO) algorithm allows for online learning of general recurrent computational graphs such as recurrent network models. It works in a streaming fashion and avoids backtracking through past activations and inputs. UORO is computationally as costly as \emph{Truncated Backpropagation Through Time} (truncated BPTT), a widespread algorithm for online learning of recurrent networks \cite{jaeger2002tutorial}. UORO is a modification of \emph{NoBackTrack} \cite{DBLP:journals/corr/OllivierC15} that bypasses the need for model sparsity and makes implementation easy in current deep learning frameworks, even for complex models. Like NoBackTrack, UORO provides unbiased gradient estimates; unbiasedness is the core hypothesis in stochastic gradient descent theory, without which convergence to a local optimum is not guaranteed. On the contrary, truncated BPTT does not provide this property, leading to possible divergence. On synthetic tasks where truncated BPTT is shown to diverge, U

ORO converges. For instance, when a parameter has a positive short-term but nega
tive long-term influence, truncated BPTT diverges unless the truncation span is
very significantly longer than the intrinsic temporal range of the interactions,
 while UORO performs well thanks to the unbiasedness of its gradients.

****
Training RNNs as Fast as CNNs
Tao Lei,Yu Zhang,Yoav Artzi
Common recurrent neural network architectures scale poorly due to the intrinsic
difficulty in parallelizing their state computations. In this work, we propose t
he Simple Recurrent Unit (SRU) architecture, a recurrent unit that simplifies th
e computation and exposes more parallelism. In SRU, the majority of computation
for each step is independent of the recurrence and can be easily parallelized. S
RU is as fast as a convolutional layer and 5-10x faster than an optimized LSTM i
mplementation. We study SRUs on a wide range of applications, including classif
ication, question answering, language modeling, translation and speech recogniti
on. Our experiments demonstrate the effectiveness of SRU and the trade-off it en
ables between speed and performance.
****
Learning Wasserstein Embeddings
Nicolas Courty,Rémi Flamary,Mélanie Ducoffe
The Wasserstein distance received a lot of attention recently in the community o
f machine learning, especially for its principled way of comparing distributions
. It has found numerous applications in several hard problems, such as domain ad
aptation, dimensionality reduction or generative models. However, its use is sti
ll limited by a heavy computational cost. Our goal is to alleviate this problem
by providing an approximation mechanism that allows to break its inherent comple
xity. It relies on the search of an embedding where the Euclidean distance mimic
s the Wasserstein distance. We show that such an embedding can be found with a s
iamese architecture associated with a decoder network that allows to move from t
he embedding space back to the original input space. Once this embedding has bee
n found, computing optimization problems in the Wasserstein space (e.g. barycent
ers, principal directions or even archetypes) can be conducted extremely fast. N
umerical experiments supporting this idea are conducted on image datasets, and s
how the wide potential benefits of our method.
****
Quantitatively Evaluating GANs With Divergences Proposed for Training
Daniel Jiwoong Im,He Ma,Graham W. Taylor,Kristin Branson
Generative adversarial networks (GANs) have been extremely effective in approxim
ating complex distributions of high-dimensional, input data samples, and substan
tial progress has been made in understanding and improving GAN performance in te
rms of both theory and application.
However, we currently lack quantitative methods for model assessment. Because of
 this, while many GAN variants being proposed, we have relatively little underst
anding of their relative abilities. In this paper, we evaluate the performance o
f various types of GANs using divergence and distance functions typically used o
nly for training. We observe consistency across the various proposed metrics and
, interestingly, the test-time metrics do not favour networks that use the same
training-time criterion. We also compare the proposed metrics to human perceptua
l scores.
****
Neural Networks with Block Diagonal Inner Product Layers
Amy Nesky,Quentin Stout
Artificial neural networks have opened up a world of possibilities in data scien
ce and artificial intelligence, but neural networks are cumbersome tools that gr
ow with the complexity of the learning problem. We make contributions to this is
sue by considering a modified version of the fully connected layer we call a blo
ck diagonal inner product layer. These modified layers have weight matrices that
 are block diagonal, turning a single fully connected layer into a set of densel
y connected neuron groups. This idea is a natural extension of group, or depthwi

se separable, convolutional layers applied to the fully connected layers. Block
diagonal inner product layers can be achieved by either initializing a purely bl
ock diagonal weight matrix or by iteratively pruning off diagonal block entries.
 This method condenses network storage and speeds up the run time without signif
icant adverse effect on the testing accuracy, thus offering a new approach to im
prove network computation efficiency.
****
AUTOMATA GUIDED HIERARCHICAL REINFORCEMENT LEARNING FOR ZERO-SHOT SKILL COMPOSIT
ION
Xiao Li,Yao Ma,Calin Belta
An obstacle that prevents the wide adoption of (deep) reinforcement learning (RL
) in control systems is its need for a large number of interactions with the env
ironment in order to master a skill. The learned skill usually generalizes poorl
y across domains and re-training is often necessary when presented with a new ta
sk. We present a framework that combines techniques in \textit{formal methods} w
ith \textit{hierarchical reinforcement learning} (HRL). The set of techniques we
 provide allows for the convenient specification of tasks with logical expressio
ns, learns hierarchical policies (meta-controller and low-level controllers) wit
h well-defined intrinsic rewards using any RL methods and is able to construct n
ew skills from existing ones without additional learning. We evaluate the propos
ed methods in a simple grid world simulation as well as simulation on a Baxter r
obot.
****
DNA-GAN: Learning Disentangled Representations from Multi-Attribute Images
Taihong Xiao,Jiapeng Hong,Jinwen Ma
Disentangling factors of variation has always been a challenging problem in repr
esentation learning. Existing algorithms suffer from many limitations, such as u
npredictable disentangling factors, bad quality of generated images from encodin
gs, lack of identity information, etc. In this paper, we proposed a supervised a
lgorithm called DNA-GAN trying to disentangle different attributes of images. Th
e latent representations of images are DNA-like, in which each individual piece
represents an independent factor of variation. By annihilating the recessive pie
ce and swapping a certain piece of two latent representations, we obtain another
 two different representations which could be decoded into images. In order to o
btain realistic images and also disentangled representations, we introduced the
discriminator for adversarial training. Experiments on Multi-PIE and CelebA data
sets demonstrate the effectiveness of our method and the advantage of overcoming
 limitations existing in other methods.
****
On the Convergence of Adam and Beyond
Sashank J. Reddi,Satyen Kale,Sanjiv Kumar
 Several recently proposed stochastic optimization methods that have been succes
sfully used in training deep networks such as RMSProp, Adam, Adadelta, Nadam are
 based on using gradient updates scaled by square roots of exponential moving av
erages of squared past gradients. In many applications, e.g. learning with large
 output spaces, it has been empirically observed that these algorithms fail to c
onverge to an optimal solution (or a critical point in nonconvex settings). We s
how that one cause for such failures is the exponential moving average used in t
he algorithms. We provide an explicit example of a simple convex optimization se
tting where Adam does not converge to the optimal solution, and describe the pre
cise problems with the previous analysis of Adam algorithm. Our analysis suggest
s that the convergence issues can be fixed by endowing such algorithms with ``lo
ng-term memory'' of past gradients, and propose new variants of the Adam algorit
hm which not only fix the convergence issues but often also lead to improved emp
irical performance.
****
Model Specialization for Inference Via End-to-End Distillation, Pruning, and Cas
cades
Daniel Kang,Karey Shi,Thao Ngyuen,Stephanie Mallard,Peter Bailis,Matei Zaharia
The availability of general-purpose reference and benchmark datasets such as

ImageNet have spurred the development of general-purpose popular reference
model architectures and pre-trained weights. However, in practice, neural net-
works are often employed to perform specific, more restrictive tasks, that are
narrower in scope and complexity. Thus, simply fine-tuning or transfer learn-
ing from a general-purpose network inherits a large computational cost that may
not be necessary for a given task. In this work, we investigate the potential fo
r
model specialization, or reducing a model’s computational footprint by leverag-
ing task-specific knowledge, such as a restricted inference distribution. We stu
dy
three methods for model specialization—1) task-aware distillation, 2) task-aware
pruning, and 3) specialized model cascades—and evaluate their performance on
a range of classification tasks. Moreover, for the first time, we investigate ho
w
these techniques complement one another, enabling up to 5× speedups with no
loss in accuracy and 9.8× speedups while remaining within 2.5% of a highly ac-
curate ResNet on specialized image classification tasks. These results suggest t
hat
simple and easy-to-implement specialization procedures may benefit a large num-
ber practical applications in which the representational power of general-purpos
e
networks need not be inherited.
****
Hallucinating brains with artificial brains
Peiye Zhuang,Alexander G. Schwing,Oluwasanmi Koyejo
Human brain function as measured by functional magnetic resonance imaging
(fMRI), exhibits a rich diversity. In response, understanding the individual var
iability
of brain function and its association with behavior has become one of the
major concerns in modern cognitive neuroscience. Our work is motivated by the
view that generative models provide a useful tool for understanding this variabi
lity.
To this end, this manuscript presents two novel generative models trained
on real neuroimaging data which synthesize task-dependent functional brain image
s.
Brain images are high dimensional tensors which exhibit structured spatial
correlations. Thus, both models are 3D conditional Generative Adversarial networ
ks
(GANs) which apply Convolutional Neural Networks (CNNs) to learn an
abstraction of brain image representations. Our results show that the generated
brain images are diverse, yet task dependent. In addition to qualitative evaluat
ion,
we utilize the generated synthetic brain volumes as additional training data to
improve
downstream fMRI classifiers (also known as decoding, or brain reading).
Our approach achieves significant improvements for a variety of datasets, classi
fi-
cation tasks and evaluation scores. Our classification results provide a quantit
ative
evaluation of the quality of the generated images, and also serve as an addition
al
contribution of this manuscript.
****
Training Confidence-calibrated Classifiers for Detecting Out-of-Distribution Sam
ples
Kimin Lee,Honglak Lee,Kibok Lee,Jinwoo Shin
The problem of detecting whether a test sample is from in-distribution (i.e., tr
aining distribution by a classifier) or out-of-distribution sufficiently differe
nt from it arises in many real-world machine learning applications. However, the
 state-of-art deep neural networks are known to be highly overconfident in their

predictions, i.e., do not distinguish in- and out-of-distributions. Recently, to handle this issue, several threshold-based detectors have been proposed given pre-trained neural classifiers. However, the performance of prior works highly depends on how to train the classifiers since they only focus on improving inference procedures. In this paper, we develop a novel training method for classifiers so that such inference algorithms can work better. In particular, we suggest two additional terms added to the original loss (e.g., cross entropy). The first one forces samples from out-of-distribution less confident by the classifier and the second one is for (implicitly) generating most effective training samples for the first one. In essence, our method jointly trains both classification and generative neural networks for out-of-distribution. We demonstrate its effectiveness using deep convolutional neural networks on various popular image datasets.
****
Synthesizing realistic neural population activity patterns using Generative Adversarial Networks
Manuel Molano-Mazon,Arno Onken,Eugenio Piasini*,Stefano Panzeri*
The ability to synthesize realistic patterns of neural activity is crucial for studying neural information processing. Here we used the Generative Adversarial Networks (GANs) framework to simulate the concerted activity of a population of neurons.
We adapted the Wasserstein-GAN variant to facilitate the generation of unconstrained neural population activity patterns while still benefiting from parameter sharing in the temporal domain.
We demonstrate that our proposed GAN, which we termed Spike-GAN, generates spike trains that match accurately the first- and second-order statistics of datasets of tens of neurons and also approximates well their higher-order statistics. We applied Spike-GAN to a real dataset recorded from salamander retina and showed that it performs as well as state-of-the-art approaches based on the maximum entropy and the dichotomized Gaussian frameworks. Importantly, Spike-GAN does not require to specify a priori the statistics to be matched by the model, and so constitutes a more flexible method than these alternative approaches.
Finally, we show how to exploit a trained Spike-GAN to construct 'importance maps' to detect the most relevant statistical structures present in a spike train.

Spike-GAN provides a powerful, easy-to-use technique for generating realistic spiking neural activity and for describing the most relevant features of the large-scale neural population recordings studied in modern systems neuroscience.

****
Real-valued (Medical) Time Series Generation with Recurrent Conditional GANs
Stephanie Hyland,Cristóbal Esteban,Gunnar Rätsch
Generative Adversarial Networks (GANs) have shown remarkable success as a framework for training models to produce realistic-looking data. In this work, we propose a Recurrent GAN (RGAN) and Recurrent Conditional GAN (RCGAN) to produce realistic real-valued multi-dimensional time series, with an emphasis on their application to medical data. RGANs make use of recurrent neural networks (RNNs) in the generator and the discriminator. In the case of RCGANs, both of these RNNs are conditioned on auxiliary information. We demonstrate our models in a set of toy datasets, where we show visually and quantitatively (using sample likelihood and maximum mean discrepancy) that they can successfully generate realistic time-series. We also describe novel evaluation methods for GANs, where we generate a synthetic labelled training dataset, and evaluate on a real test set the performance of a model trained on the synthetic data, and vice-versa. We illustrate with these metrics that RCGANs can generate time-series data useful for supervised training, with only minor degradation in performance on real test data. This is demonstrated on digit classification from ‘serialised’ MNIST and by training an early warning system on a medical dataset of 17,000 patients from an intensive care unit. We further discuss and analyse the privacy concerns that may arise when using RCGANs to generate realistic synthetic medical time series data, and demonstrate results from differentially private training of the RCGAN.

****
Kernel Implicit Variational Inference
Jiaxin Shi,Shengyang Sun,Jun Zhu
Recent progress in variational inference has paid much attention to the flexibility of variational posteriors. One promising direction is to use implicit distributions, i.e., distributions without tractable densities as the variational posterior. However, existing methods on implicit posteriors still face challenges of noisy estimation and computational infeasibility when applied to models with high-dimensional latent variables. In this paper, we present a new approach named Kernel Implicit Variational Inference that addresses these challenges. As far as we know, for the first time implicit variational inference is successfully applied to Bayesian neural networks, which shows promising results on both regression and classification tasks.
****
Residual Loss Prediction: Reinforcement Learning With No Incremental Feedback
Hal Daumé III,John Langford,Amr Sharaf
We consider reinforcement learning and bandit structured prediction problems with very sparse loss feedback: only at the end of an episode. We introduce a novel algorithm, RESIDUAL LOSS PREDICTION (RESLOPE), that solves such problems by automatically learning an internal representation of a denser reward function. RESLOPE operates as a reduction to contextual bandits, using its learned loss representation to solve the credit assignment problem, and a contextual bandit oracle to trade-off exploration and exploitation. RESLOPE enjoys a no-regret reduction-style theoretical guarantee and outperforms state of the art reinforcement learning algorithms in both MDP environments and bandit structured prediction settings.
****
POLICY DRIVEN GENERATIVE ADVERSARIAL NETWORKS FOR ACCENTED SPEECH GENERATION
Prannay Khosla,Preethi Jyothi,Vinay P. Namboodiri,Mukundhan Srinivasan
In this paper, we propose the generation of accented speech using generative adversarial
networks. Through this work we make two main contributions a) The
ability to condition latent representations while generating realistic speech samples
b) The ability to efficiently generate long speech samples by using a novel
latent variable transformation module that is trained using policy gradients. Previous
methods are limited in being able to generate only relatively short samples
or are not very efficient at generating long samples. The generated speech samples
are validated through a number of various evaluation measures viz, a WGAN
critic loss and through subjective scores on user evaluations against competitive
speech synthesis baselines and detailed ablation analysis of the proposed model.
The evaluations demonstrate that the model generates realistic long speech samples
conditioned on accent efficiently.
****
Discrete Autoencoders for Sequence Models
Lukasz Kaiser,Samy Bengio
Recurrent models for sequences have been recently successful at many tasks, especially for language modeling
and machine translation. Nevertheless, it remains challenging to extract good representations from
these models. For instance, even though language has a clear hierarchical structure going from characters
through words to sentences, it is not apparent in current language models.
We propose to improve the representation in sequence models by
augmenting current approaches with an autoencoder that is forced to compress
the sequence through an intermediate discrete latent space. In order to propagat

e gradients
though this discrete representation we introduce an improved semantic hashing technique.
We show that this technique performs well on a newly proposed quantitative efficiency measure.
We also analyze latent codes produced by the model showing how they correspond to
words and phrases. Finally, we present an application of the autoencoder-augmented
model to generating diverse translations.
****
Distributed Distributional Deterministic Policy Gradients
Gabriel Barth-Maron,Matthew W. Hoffman,David Budden,Will Dabney,Dan Horgan,Dhruva TB,Alistair Muldal,Nicolas Heess,Timothy Lillicrap
This work adopts the very successful distributional perspective on reinforcement learning and adapts it to the continuous control setting. We combine this within a distributed framework for off-policy learning in order to develop what we call the Distributed Distributional Deep Deterministic Policy Gradient algorithm, D4PG. We also combine this technique with a number of additional, simple improvements such as the use of N-step returns and prioritized experience replay. Experimentally we examine the contribution of each of these individual components, and show how they interact, as well as their combined contributions. Our results show that across a wide variety of simple control tasks, difficult manipulation tasks, and a set of hard obstacle-based locomotion tasks the D4PG algorithm achieves state of the art performance.
****
TOWARDS ROBOT VISION MODULE DEVELOPMENT WITH EXPERIENTIAL ROBOT LEARNING
Ahmed A Aly,Joanne Bechta Dugan
n this paper we present a thrust in three directions of visual development us- ing supervised and semi-supervised techniques. The first is an implementation of semi-supervised object detection and recognition using the principles of Soft At- tention and Generative Adversarial Networks (GANs). The second and the third are supervised networks that learn basic concepts of spatial locality and quantity respectively using Convolutional Neural Networks (CNNs). The three thrusts to- gether are based on the approach of Experiential Robot Learning, introduced in previous publication. While the results are unripe for implementation, we believe they constitute a stepping stone towards autonomous development of robotic vi- sual modules.
****
Breaking the Softmax Bottleneck: A High-Rank RNN Language Model
Zhilin Yang,Zihang Dai,Ruslan Salakhutdinov,William W. Cohen
We formulate language modeling as a matrix factorization problem, and show that the expressiveness of Softmax-based models (including the majority of neural language models) is limited by a Softmax bottleneck. Given that natural language is highly context-dependent, this further implies that in practice Softmax with distributed word embeddings does not have enough capacity to model natural language. We propose a simple and effective method to address this issue, and improve the state-of-the-art perplexities on Penn Treebank and WikiText-2 to 47.69 and 40.68 respectively. The proposed method also excels on the large-scale 1B Word dataset, outperforming the baseline by over 5.6 points in perplexity.
****
Debiasing Evidence Approximations: On Importance-weighted Autoencoders and Jackknife Variational Inference
Sebastian Nowozin
The importance-weighted autoencoder (IWAE) approach of Burda et al. defines a sequence of increasingly tighter bounds on the marginal likelihood of latent variable models. Recently, Cremer et al. reinterpreted the IWAE bounds as ordinary variational evidence lower bounds (ELBO) applied to increasingly accurate variational distributions. In this work, we provide yet another perspective on the IWAE bounds. We interpret each IWAE bound as a biased estimator of the true marginal

likelihood where for the bound defined on K samples we show the bias to be of order O(1/K). In our theoretical analysis of the IWAE objective we derive asymptotic bias and variance expressions. Based on this analysis we develop jackknife variational inference (JVI),
a family of bias-reduced estimators reducing the bias to $O(K^{-(m+1)})$ for any given m < K while retaining computational efficiency. Finally, we demonstrate that JVI leads to improved evidence estimates in variational autoencoders. We also report first results on applying JVI to learning variational autoencoders.

Our implementation is available at https://github.com/Microsoft/jackknife-variational-inference
****
Maintaining cooperation in complex social dilemmas using deep reinforcement learning
Alexander Peysakhovich,Adam Lerer
Social dilemmas are situations where individuals face a temptation to increase their payoffs at a cost to total welfare. Building artificially intelligent agents that achieve good outcomes in these situations is important because many real world interactions include a tension between selfish interests and the welfare of others. We show how to modify modern reinforcement learning methods to construct agents that act in ways that are simple to understand, nice (begin by cooperating), provokable (try to avoid being exploited), and forgiving (try to return to mutual cooperation). We show both theoretically and experimentally that such agents can maintain cooperation in Markov social dilemmas. Our construction does not require training methods beyond a modification of self-play, thus if an environment is such that good strategies can be constructed in the zero-sum case (eg. Atari) then we can construct agents that solve social dilemmas in this environment.
****
SHADE: SHAnnon DEcay Information-Based Regularization for Deep Learning
Michael Blot,Thomas Robert,Nicolas Thome,Matthieu Cord
Regularization is a big issue for training deep neural networks. In this paper, we propose a new information-theory-based regularization scheme named SHADE for SHAnnon DEcay. The originality of the approach is to define a prior based on conditional entropy, which explicitly decouples the learning of invariant representations in the regularizer and the learning of correlations between inputs and labels in the data fitting term. We explain why this quantity makes our model able to achieve invariance with respect to input variations. We empirically validate the efficiency of our approach to improve classification performances compared to standard regularization schemes on several standard architectures.
****
Towards Reverse-Engineering Black-Box Neural Networks
Seong Joon Oh,Max Augustin,Mario Fritz,Bernt Schiele
Many deployed learned models are black boxes: given input, returns output. Internal information about the model, such as the architecture, optimisation procedure, or training data, is not disclosed explicitly as it might contain proprietary information or make the system more vulnerable. This work shows that such attributes of neural networks can be exposed from a sequence of queries. This has multiple implications. On the one hand, our work exposes the vulnerability of black-box neural networks to different types of attacks -- we show that the revealed internal information helps generate more effective adversarial examples against the black box model. On the other hand, this technique can be used for better protection of private content from automatic recognition models using adversarial examples. Our paper suggests that it is actually hard to draw a line between white box and black box models.
****
Go for a Walk and Arrive at the Answer: Reasoning Over Paths in Knowledge Bases using Reinforcement Learning
Rajarshi Das,Shehzaad Dhuliawala,Manzil Zaheer,Luke Vilnis,Ishan Durugkar,Akshay Krishnamurthy,Alex Smola,Andrew McCallum

Knowledge bases (KB), both automatically and manually constructed, are often incomplete --- many valid facts can be inferred from the KB by synthesizing existing information. A popular approach to KB completion is to infer new relations by combinatory reasoning over the information found along other paths connecting a pair of entities. Given the enormous size of KBs and the exponential number of paths, previous path-based models have considered only the problem of predicting a missing relation given two entities, or evaluating the truth of a proposed triple. Additionally, these methods have traditionally used random paths between fixed entity pairs or more recently learned to pick paths between them. We propose a new algorithm, MINERVA, which addresses the much more difficult and practical task of answering questions where the relation is known, but only one entity. Since random walks are impractical in a setting with unknown destination and combinatorially many paths from a start node, we present a neural reinforcement learning approach which learns how to navigate the graph conditioned on the input query to find predictive paths. On a comprehensive evaluation on seven knowledge base datasets, we found MINERVA to be competitive with many current state-of-the-art methods.
****

Learning to Encode Text as Human-Readable Summaries using Generative Adversarial Networks
Yau-Shian Wang,Hung-Yi Lee
Auto-encoders compress input data into a latent-space representation and reconstruct the original data from the representation. This latent representation is not easily interpreted by humans. In this paper, we propose training an auto-encoder that encodes input text into human-readable sentences. The auto-encoder is composed of a generator and a reconstructor. The generator encodes the input text into a shorter word sequence, and the reconstructor recovers the generator input from the generator output.
To make the generator output human-readable, a discriminator restricts the output of the generator to resemble human-written sentences. By taking the generator output as the summary of the input text, abstractive summarization is achieved without document-summary pairs as training data. Promising results are shown on both English and Chinese corpora.
****

Learnability of Learned Neural Networks
Rahul Anand Sharma,Navin Goyal,Monojit Choudhury,Praneeth Netrapalli
This paper explores the simplicity of learned neural networks under various settings: learned on real vs random data, varying size/architecture and using large minibatch size vs small minibatch size. The notion of simplicity used here is that of learnability i.e., how accurately can the prediction function of a neural network be learned from labeled samples from it. While learnability is different from (in fact often higher than) test accuracy, the results herein suggest that there is a strong correlation between small generalization errors and high learnability.
This work also shows that there exist significant qualitative differences in shallow networks as compared to popular deep networks. More broadly, this paper extends in a new direction, previous work on understanding the properties of learned neural networks. Our hope is that such an empirical study of understanding learned neural networks might shed light on the right assumptions that can be made for a theoretical study of deep learning.
****

One-shot and few-shot learning of word embeddings
Andrew Kyle Lampinen,James Lloyd McClelland
Standard deep learning systems require thousands or millions of examples to learn a concept, and cannot integrate new concepts easily. By contrast, humans have an incredible ability to do one-shot or few-shot learning. For instance, from just hearing a word used in a sentence, humans can infer a great deal about it, by leveraging what the syntax and semantics of the surrounding words tells us. Here, we draw inspiration from this to highlight a simple technique by which deep recurrent networks can similarly exploit their prior knowledge to learn a useful

representation for a new word from little data. This could make natural language processing systems much more flexible, by allowing them to learn continually from the new words they encounter.
****

Learning to Write by Learning the Objective
Ari Holtzman,Jan Buys,Maxwell Forbes,Antoine Bosselut,Yejin Choi
Recurrent Neural Networks (RNNs) are powerful autoregressive sequence models for learning prevalent patterns in natural language. Yet language generated by RNNs often shows several degenerate characteristics that are uncommon in human language; while fluent, RNN language production can be overly generic, repetitive, and even self-contradictory. We postulate that the objective function optimized by RNN language models, which amounts to the overall perplexity of a text, is not expressive enough to capture the abstract qualities of good generation such as Grice’s Maxims. In this paper, we introduce a general learning framework that can construct a decoding objective better suited for generation. Starting with a generatively trained RNN language model, our framework learns to construct a substantially stronger generator by combining several discriminatively trained models that can collectively address the limitations of RNN generation. Human evaluation demonstrates that text generated by the resulting generator is preferred over that of baselines by a large margin and significantly enhances the overall coherence, style, and information content of the generated text.
****

Building Generalizable Agents with a Realistic and Rich 3D Environment
Yi Wu,Yuxin Wu,Georgia Gkioxari,Yuandong Tian
Teaching an agent to navigate in an unseen 3D environment is a challenging task, even in the event of simulated environments. To generalize to unseen environments, an agent needs to be robust to low-level variations (e.g. color, texture, object changes), and also high-level variations (e.g. layout changes of the environment). To improve overall generalization, all types of variations in the environment have to be taken under consideration via different level of data augmentation steps. To this end, we propose House3D, a rich, extensible and efficient environment that contains 45,622 human-designed 3D scenes of visually realistic houses, ranging from single-room studios to multi-storied houses, equipped with a diverse set of fully labeled 3D objects, textures and scene layouts, based on the SUNCG dataset (Song et al., 2017). The diversity in House3D opens the door towards scene-level augmentation, while the label-rich nature of House3D enables us to inject pixel- & task-level augmentations such as domain randomization (Tobin et al., 2017) and multi-task training. Using a subset of houses in House3D, we show that reinforcement learning agents trained with an enhancement of different levels of augmentations perform much better in unseen environments than our baselines with raw RGB input by over 8% in terms of navigation success rate. House3D is publicly available at http://github.com/facebookresearch/House3D.
****

Generative Models for Alignment and Data Efficiency in Language
Dustin Tran,Yura Burda,Ilya Sutskever
We examine how learning from unaligned data can improve both the data efficiency of supervised tasks as well as enable alignments without any supervision. For example, consider unsupervised machine translation: the input is two corpora of English and French, and the task is to translate from one language to the other but without any pairs of English and French sentences. To address this, we develop feature-matching autoencoders (FMAEs). FMAEs ensure that the marginal distribution of feature layers are preserved across forward and inverse mappings between domains. We show that FMAEs achieve state of the art for data efficiency and alignment across three tasks: text decipherment, sentiment transfer, and neural machine translation for English-to-German and English-to-French. Most compellingly, FMAEs achieve state of the art for neural translation with limited supervision, with significant BLEU score differences of up to 5.7 and 6.3 over traditional supervised models. Furthermore, on English-to-German, they outperform last year's best fully supervised models such as ByteNet (Kalchbrenner et al., 2016) while using only half as many supervised examples.

****
Interpreting Deep Classification Models With Bayesian Inference
Hanshu Yan,Jiashi Feng
In this paper, we propose a novel approach to interpret a well-trained classific ation model through systematically investigating effects of its hidden units on prediction making. We search for the core hidden units responsible for predictin g inputs as the class of interest under the generative Bayesian inference framew ork. We model such a process of unit selection as an Indian Buffet Process, and derive a simplified objective function via the MAP asymptotic technique. The ind uced binary optimization problem is efficiently solved with a continuous relaxat ion method by attaching a Switch Gate layer to the hidden layers of interest. Th e resulted interpreter model is thus end-to-end optimized via standard gradient back-propagation. Experiments are conducted with two popular deep convolutional classifiers, respectively well-trained on the MNIST dataset and the CI- FAR10 da taset. The results demonstrate that the proposed interpreter successfully finds the core hidden units most responsible for prediction making. The modified model , only with the selected units activated, can hold correct predictions at a high rate. Besides, this interpreter model is also able to extract the most informat ive pixels in the images by connecting a Switch Gate layer to the input layer.

****
Deep Function Machines: Generalized Neural Networks for Topological Layer Expres sion
William H. Guss
In this paper we propose a generalization of deep neural networks called deep fu nction machines (DFMs). DFMs act on vector spaces of arbitrary (possibly infinit e) dimension and we show that a family of DFMs are invariant to the dimension of input data; that is, the parameterization of the model does not directly hinge on the quality of the input (eg. high resolution images). Using this generaliza tion we provide a new theory of universal approximation of bounded non-linear op erators between function spaces. We then suggest that DFMs provide an expressive framework for designing new neural network layer types with topological conside rations in mind. Finally, we introduce a novel architecture, RippLeNet, for reso lution invariant computer vision, which empirically achieves state of the art in variance.
****
Unseen Class Discovery in Open-world Classification
Lei Shu,Hu Xu,Bing Liu
This paper concerns open-world classification, where the classifier not only nee ds to classify test examples into seen classes that have appeared in training bu t also reject examples from unseen or novel classes that have not appeared in tr aining. Specifically, this paper focuses on discovering the hidden unseen classe s of the rejected examples. Clearly, without prior knowledge this is difficult. However, we do have the data from the seen training classes, which can tell us w hat kind of similarity/difference is expected for examples from the same class o r from different classes. It is reasonable to assume that this knowledge can be transferred to the rejected examples and used to discover the hidden unseen clas ses in them. This paper aims to solve this problem. It first proposes a joint op en classification model with a sub-model for classifying whether a pair of examp les belongs to the same or different classes. This sub-model can serve as a dist ance function for clustering to discover the hidden classes of the rejected exam ples. Experimental results show that the proposed model is highly promising.

****
Boosting the Actor with Dual Critic
Bo Dai,Albert Shaw,Niao He,Lihong Li,Le Song
This paper proposes a new actor-critic-style algorithm called Dual Actor-Critic or Dual-AC. It is derived in a principled way from the Lagrangian dual form of the Bellman optimality equation, which can be viewed as a two-player game betwee n the actor and a critic-like function, which is named as dual critic. Compared

to its actor-critic relatives, Dual-AC has the desired property that the actor and dual critic are updated cooperatively to optimize the same objective functio n, providing a more transparent way for learning the critic that is directly rel ated to the objective function of the actor. We then provide a concrete algorith m that can effectively solve the minimax optimization problem, using techniques of multi-step bootstrapping, path regularization, and stochastic dual ascent alg orithm. We demonstrate that the proposed algorithm achieves the state-of-the-art performances across several benchmarks.
****
Mastering the Dungeon: Grounded Language Learning by Mechanical Turker Descent
Zhilin Yang,Saizheng Zhang,Jack Urbanek,Will Feng,Alexander Miller,Arthur Szlam, Douwe Kiela,Jason Weston
Contrary to most natural language processing research, which makes use of static datasets, humans learn language interactively, grounded in an environment. In this work we propose an interactive learning procedure called Mechanical Turker Descent (MTD) that trains agents to execute natural language commands grounded i n a fantasy text adventure game. In MTD, Turkers compete to train better agents in the short term, and collaborate by sharing their agents' skills in the long t erm. This results in a gamified, engaging experience for the Turkers and a bette r quality teaching signal for the agents compared to static datasets, as the Tur kers naturally adapt the training data to the agent's abilities.
****
Deep Asymmetric Multi-task Feature Learning
Hae Beom Lee,Eunho Yang,Sung Ju Hwang
We propose Deep Asymmetric Multitask Feature Learning (Deep-AMTFL) which can lea rn deep representations shared across multiple tasks while effectively preventin g negative transfer that may happen in the feature sharing process. Specifically , we introduce an asymmetric autoencoder term that allows reliable predictors fo r the easy tasks to have high contribution to the feature learning while suppres sing the influences of unreliable predictors for more difficult tasks. This allo ws the learning of less noisy representations, and enables unreliable predictors to exploit knowledge from the reliable predictors via the shared latent feature s. Such asymmetric knowledge transfer through shared features is also more scala ble and efficient than inter-task asymmetric transfer. We validate our Deep-AMTF L model on multiple benchmark datasets for multitask learning and image classifi cation, on which it significantly outperforms existing symmetric and asymmetric multitask learning models, by effectively preventing negative transfer in deep f eature learning.
****
Policy Optimization by Genetic Distillation
Tanmay Gangwani,Jian Peng
Genetic algorithms have been widely used in many practical optimization problems
.
Inspired by natural selection, operators, including mutation, crossover
and selection, provide effective heuristics for search and black-box optimizatio
n.
However, they have not been shown useful for deep reinforcement learning, possib
ly
due to the catastrophic consequence of parameter crossovers of neural networks.
Here, we present Genetic Policy Optimization (GPO), a new genetic algorithm
for sample-efficient deep policy optimization. GPO uses imitation learning
for policy crossover in the state space and applies policy gradient methods for
mutation.
Our experiments on MuJoCo tasks show that GPO as a genetic algorithm
is able to provide superior performance over the state-of-the-art policy gradien
t
methods and achieves comparable or higher sample efficiency.
****
Learning to Infer Graphics Programs from Hand-Drawn Images
Kevin Ellis,Daniel Ritchie,Armando Solar-Lezama,Joshua B. Tenenbaum

We introduce a model that learns to convert simple hand drawings
into graphics programs written in a subset of \LaTeX.~The model
combines techniques from deep learning and program synthesis. We
learn a convolutional neural network that proposes plausible drawing
primitives that explain an image. These drawing primitives are like
a trace of the set of primitive commands issued by a graphics
program. We learn a model that uses program synthesis techniques to
recover a graphics program from that trace. These programs have
constructs like variable bindings, iterative loops, or simple kinds
of conditionals. With a graphics program in hand, we can correct
errors made by the deep network and extrapolate drawings. Taken
together these results are a step towards agents that induce useful,
human-readable programs from perceptual input.

****

Continuous-Time Flows for Efficient Inference and Density Estimation

Changyou Chen,Chunyuan Li,Liqun Chen,Wenlin Wang,Yunchen Pu,Lawrence Carin

Two fundamental problems in unsupervised learning are efficient inference for latent-variable models and robust density estimation based on large amounts of unlabeled data. For efficient inference, normalizing flows have been recently developed to approximate a target distribution arbitrarily well. In practice, however, normalizing flows only consist of a finite number of deterministic transformations, and thus they possess no guarantee on the approximation accuracy. For density estimation, the generative adversarial network (GAN) has been advanced as an appealing model, due to its often excellent performance in generating samples. In this paper, we propose the concept of {\em continuous-time flows} (CTFs), a family of diffusion-based methods that are able to asymptotically approach a target distribution. Distinct from normalizing flows and GANs, CTFs can be adopted to achieve the above two goals in one framework, with theoretical guarantees. Our framework includes distilling knowledge from a CTF for efficient inference, and learning an explicit energy-based distribution with CTFs for density estimation. Experiments on various tasks demonstrate promising performance of the proposed CTF framework, compared to related techniques.

****

Challenges in Disentangling Independent Factors of Variation

Attila Szabo,Qiyang Hu,Tiziano Portenier,Matthias Zwicker,Paolo Favaro

We study the problem of building models that disentangle independent factors of variation. Such models encode features that can efficiently be used for classification and to transfer attributes between different images in image synthesis. As data we use a weakly labeled training set, where labels indicate what single factor has changed between two data samples, although the relative value of the change is unknown. This labeling is of particular interest as it may be readily available without annotation costs. We introduce an autoencoder model and train it through constraints on image pairs and triplets. We show the role of feature dimensionality and adversarial training theoretically and experimentally. We formally prove the existence of the reference ambiguity, which is inherently present in the disentangling task when weakly labeled data is used. The numerical value of a factor has different meaning in different reference frames. When the reference depends on other factors, transferring that factor becomes ambiguous. We demonstrate experimentally that the proposed model can successfully transfer attributes on several datasets, but show also cases when the reference ambiguity occurs.

****

Deep Neural Networks as Gaussian Processes

Jaehoon Lee,Yasaman Bahri,Roman Novak,Samuel S. Schoenholz,Jeffrey Pennington,Jascha Sohl-Dickstein

It has long been known that a single-layer fully-connected neural network with an i.i.d. prior over its parameters is equivalent to a Gaussian process (GP), in the limit of infinite network width. This correspondence enables exact Bayesian inference for infinite width neural networks on regression tasks by means of evaluating the corresponding GP. Recently, kernel functions which mimic multi-layer random neural networks have been developed, but only outside of a Bayesian framework. As such, previous work has not identified that these kernels can be used as covariance functions for GPs and allow fully Bayesian prediction with a deep neural network.

In this work, we derive the exact equivalence between infinitely wide, deep, networks and GPs with a particular covariance function. We further develop a computationally efficient pipeline to compute this covariance function. We then use the resulting GP to perform Bayesian inference for deep neural networks on MNIST and CIFAR-10. We observe that the trained neural network accuracy approaches that of the corresponding GP with increasing layer width, and that the GP uncertainty is strongly correlated with trained network prediction error. We further find that test performance increases as finite-width trained networks are made wider and more similar to a GP, and that the GP-based predictions typically outperform those of finite-width networks. Finally we connect the prior distribution over weights and variances in our GP formulation to the recent development of signal propagation in random neural networks.
****

LatentPoison -- Adversarial Attacks On The Latent Space
Antonia Creswell,Biswa Sengupta,Anil A. Bharath
Robustness and security of machine learning (ML) systems are intertwined, wherein a non-robust ML system (classifiers, regressors, etc.) can be subject to attacks using a wide variety of exploits. With the advent of scalable deep learning methodologies, a lot of emphasis has been put on the robustness of supervised, unsupervised and reinforcement learning algorithms. Here, we study the robustness of the latent space of a deep variational autoencoder (dVAE), an unsupervised generative framework, to show that it is indeed possible to perturb the latent space, flip the class predictions and keep the classification probability approximately equal before and after an attack. This means that an agent that looks at the outputs of a decoder would remain oblivious to an attack.
****

META LEARNING SHARED HIERARCHIES
Kevin Frans,Jonathan Ho,Xi Chen,Pieter Abbeel,John Schulman
We develop a metalearning approach for learning hierarchically structured poli- cies, improving sample efficiency on unseen tasks through the use of shared primitives—policies that are executed for large numbers of timesteps. Specifi- cally, a set of primitives are shared within a distribution of tasks, and are switched between by task-specific policies. We provide a concrete metric for measuring the strength of such hierarchies, leading to an optimization problem for quickly reaching high reward on unseen tasks. We then present an algorithm to solve this problem end-to-end through the use of any off-the-shelf reinforcement learning method, by repeatedly sampling new tasks and resetting task-specific policies. We successfully discover meaningful motor primitives for the directional movement of four-legged robots, solely by interacting with distributions of mazes. We also demonstrate the transferability of primitives to solve long-timescale sparse-reward obstacle courses, and we enable 3D humanoid robots to robustly walk and crawl with the same policy.
****

PixelDefend: Leveraging Generative Models to Understand and Defend against Adversarial Examples
Yang Song,Taesup Kim,Sebastian Nowozin,Stefano Ermon,Nate Kushman
Adversarial perturbations of normal images are usually imperceptible to humans, but they can seriously confuse state-of-the-art machine learning models. What makes them so special in the eyes of image classifiers? In this paper, we show empirically that adversarial examples mainly lie in the low probability regions of the training distribution, regardless of attack types and targeted models. Using statistical hypothesis testing, we find that modern neural density models are surprisingly good at detecting imperceptible image perturbations. Based on this d

iscovery, we devised PixelDefend, a new approach that purifies a maliciously perturbed image by moving it back towards the distribution seen in the training data. The purified image is then run through an unmodified classifier, making our method agnostic to both the classifier and the attacking method. As a result, PixelDefend can be used to protect already deployed models and be combined with other model-specific defenses. Experiments show that our method greatly improves resilience across a wide variety of state-of-the-art attacking methods, increasing accuracy on the strongest attack from 63% to 84% for Fashion MNIST and from 32% to 70% for CIFAR-10.
****
Hierarchical Adversarially Learned Inference
Mohamed Ishmael Belghazi,Sai Rajeswar,Olivier Mastropietro,Negar Rostamzadeh,Jovana Mitrovic,Aaron Courville
We propose a novel hierarchical generative model with a simple Markovian structure and a corresponding inference model. Both the generative and inference model are trained using the adversarial learning paradigm. We demonstrate that the hierarchical structure supports the learning of progressively more abstract representations as well as providing semantically meaningful reconstructions with different levels of fidelity. Furthermore, we show that minimizing the Jensen-Shanon divergence between the generative and inference network is enough to minimize the reconstruction error. The resulting semantically meaningful hierarchical latent structure discovery is exemplified on the CelebA dataset. There, we show that the features learned by our model in an unsupervised way outperform the best handcrafted features. Furthermore, the extracted features remain competitive when compared to several recent deep supervised approaches on an attribute prediction task on CelebA. Finally, we leverage the model's inference network to achieve state-of-the-art performance on a semi-supervised variant of the MNIST digit classification task.
****
Cheap DNN Pruning with Performance Guarantees
Konstantinos Pitas,Mike Davies,Pierre Vandergheynst
Recent DNN pruning algorithms have succeeded in reducing the number of parameters in fully connected layers often with little or no drop in classification accuracy. However most of the existing pruning schemes either have to be applied during training or require a costly retraining procedure after pruning to regain classification accuracy. In this paper we propose a cheap pruning algorithm based on difference of convex (DC) optimisation. We also provide theoretical analysis for the growth in the Generalisation Error (GE) of the new pruned network. Our method can be used with any convex regulariser and allows for a controlled degradation in classification accuracy while being orders of magnitude faster than competing approaches. Experiments on common feedforward neural networks show that for sparsity levels above 90% our method achieves 10% higher classification accuracy compared to Hard Thresholding.
****
Ensemble Adversarial Training: Attacks and Defenses
Florian Tramèr,Alexey Kurakin,Nicolas Papernot,Ian Goodfellow,Dan Boneh,Patrick McDaniel
Adversarial examples are perturbed inputs designed to fool machine learning models. Adversarial training injects such examples into training data to increase robustness. To scale this technique to large datasets, perturbations are crafted using fast single-step methods that maximize a linear approximation of the model's loss.
We show that this form of adversarial training converges to a degenerate global minimum, wherein small curvature artifacts near the data points obfuscate a linear approximation of the loss. The model thus learns to generate weak perturbations, rather than defend against strong ones. As a result, we find that adversarial training remains vulnerable to black-box attacks, where we transfer perturbations computed on undefended models, as well as to a powerful novel single-step attack that escapes the non-smooth vicinity of the input data via a small random step.

We further introduce Ensemble Adversarial Training, a technique that augments tr
aining data with perturbations transferred from other models. On ImageNet, Ensem
ble Adversarial Training yields models with strong robustness to black-box attac
ks. In particular, our most robust model won the first round of the NIPS 2017 co
mpetition on Defenses against Adversarial Attacks.
****

DCN+: Mixed Objective And Deep Residual Coattention for Question Answering
Caiming Xiong,Victor Zhong,Richard Socher
Traditional models for question answering optimize using cross entropy loss, whi
ch encourages exact answers at the cost of penalizing nearby or overlapping answ
ers that are sometimes equally accurate. We propose a mixed objective that combi
nes cross entropy loss with self-critical policy learning, using rewards derived
 from word overlap to solve the misalignment between evaluation metric and optim
ization objective. In addition to the mixed objective, we introduce a deep resid
ual coattention encoder that is inspired by recent work in deep self-attention a
nd residual networks. Our proposals improve model performance across question ty
pes and input lengths, especially for long questions that requires the ability t
o capture long-term dependencies. On the Stanford Question Answering Dataset, ou
r model achieves state of the art results with 75.1% exact match accuracy and 83
.1% F1, while the ensemble obtains 78.9% exact match accuracy and 86.0% F1.
****

All-but-the-Top: Simple and Effective Postprocessing for Word Representations
Jiaqi Mu,Pramod Viswanath
Real-valued word representations have transformed NLP applications; popular exam
ples are word2vec and GloVe, recognized for their ability to capture linguistic
regularities. In this paper, we demonstrate a {\em very simple}, and yet counter
-intuitive, postprocessing technique -- eliminate the common mean vector and a f
ew top dominating directions from the word vectors -- that renders off-the-shelf
 representations {\em even stronger}. The postprocessing is empirically validate
d on a variety of lexical-level intrinsic tasks (word similarity, concept catego
rization, word analogy) and sentence-level tasks (semantic textural similarity a
nd text classification) on multiple datasets and with a variety of representatio
n methods and hyperparameter choices in multiple languages; in each case, the pr
ocessed representations are consistently better than the original ones.
****

Deep Rewiring: Training very sparse deep networks
Guillaume Bellec,David Kappel,Wolfgang Maass,Robert Legenstein
Neuromorphic hardware tends to pose limits on the connectivity of deep networks
that one can run on them. But also generic hardware and software implementations
 of deep learning run more efficiently for sparse networks. Several methods exis
t for pruning connections of a neural network after it was trained without conne
ctivity constraints. We present an algorithm, DEEP R, that enables us to train d
irectly a sparsely connected neural network. DEEP R automatically rewires the ne
twork during supervised training so that connections are there where they are mo
st needed for the task, while its total number is all the time strictly bounded.
 We demonstrate that DEEP R can be used to train very sparse feedforward and rec
urrent neural networks on standard benchmark tasks with just a minor loss in per
formance. DEEP R is based on a rigorous theoretical foundation that views rewiri
ng as stochastic sampling of network configurations from a posterior.
****

Communication Algorithms via Deep Learning
Hyeji Kim,Yihan Jiang,Ranvir B. Rana,Sreeram Kannan,Sewoong Oh,Pramod Viswanath
Coding theory is a central discipline underpinning wireline and wireless modems
that are the workhorses of the information age. Progress in coding theory is lar
gely driven by individual human ingenuity with sporadic breakthroughs over the p
ast century. In this paper we study whether it is possible to automate the disco
very of decoding algorithms via deep learning. We study a family of sequential c
odes parametrized by recurrent neural network (RNN) architectures. We show that
cre- atively designed and trained RNN architectures can decode well known sequen
tial codes such as the convolutional and turbo codes with close to optimal perfo

rmance on the additive white Gaussian noise (AWGN) channel, which itself is achi eved by breakthrough algorithms of our times (Viterbi and BCJR decoders, represe nting dynamic programing and forward-backward algorithms). We show strong gen- e ralizations, i.e., we train at a specific signal to noise ratio and block length but test at a wide range of these quantities, as well as robustness and adaptiv ity to deviations from the AWGN setting.
****
Faster Discovery of Neural Architectures by Searching for Paths in a Large Model
Hieu Pham,Melody Y. Guan,Barret Zoph,Quoc V. Le,Jeff Dean
We propose Efficient Neural Architecture Search (ENAS), a faster and less expens ive approach to automated model design than previous methods. In ENAS, a control ler learns to discover neural network architectures by searching for an optimal path within a larger model. The controller is trained with policy gradient to se lect a path that maximizes the expected reward on the validation set. Meanwhile the model corresponding to the selected path is trained to minimize the cross en tropy loss. On the Penn Treebank dataset, ENAS can discover a novel architecture thats achieves a test perplexity of 57.8, which is state-of-the-art among autom atic model design methods on Penn Treebank. On the CIFAR-10 dataset, ENAS can de sign novel architectures that achieve a test error of 2.89%, close to the 2.65% achieved by standard NAS (Zoph et al., 2017). Most importantly, our experiments show that ENAS is more than 10x faster and 100x less resource-demanding than NAS .
****
Data-efficient Deep Reinforcement Learning for Dexterous Manipulation
Ivo Popov,Nicolas Heess,Timothy P. Lillicrap,Roland Hafner,Gabriel Barth-Maron,M atej Vecerik,Thomas Lampe,Tom Erez,Yuval Tassa,Martin Riedmiller
Grasping an object and precisely stacking it on another is a difficult task for traditional robotic control or hand-engineered approaches. Here we examine the p roblem in simulation and provide techniques aimed at solving it via deep reinfor cement learning. We introduce two straightforward extensions to the Deep Determi nistic Policy Gradient algorithm (DDPG), which make it significantly more data-e fficient and scalable. Our results show that by making extensive use of off-poli cy data and replay, it is possible to find high-performance control policies. Fu rther, our results hint that it may soon be feasible to train successful stackin g policies by collecting interactions on real robots.
****
Demystifying MMD GANs
Miko■aj Bi■kowski,Danica J. Sutherland,Michael Arbel,Arthur Gretton
We investigate the training and performance of generative adversarial networks u sing the Maximum Mean Discrepancy (MMD) as critic, termed MMD GANs. As our main theoretical contribution, we clarify the situation with bias in GAN loss functio ns raised by recent work: we show that gradient estimators used in the optimizat ion process for both MMD GANs and Wasserstein GANs are unbiased, but learning a discriminator based on samples leads to biased gradients for the generator param eters. We also discuss the issue of kernel choice for the MMD critic, and charac terize the kernel corresponding to the energy distance used for the Cramér GAN c ritic. Being an integral probability metric, the MMD benefits from training stra tegies recently developed for Wasserstein GANs. In experiments, the MMD GAN is a ble to employ a smaller critic network than the Wasserstein GAN, resulting in a simpler and faster-training algorithm with matching performance. We also propose an improved measure of GAN convergence, the Kernel Inception Distance, and show how to use it to dynamically adapt learning rates during GAN training.
****
Diffusing Policies : Towards Wasserstein Policy Gradient Flows
Pierre H. Richemond,Brendan Maginnis
Policy gradients methods often achieve better performance when the change in pol icy is limited to a small Kullback-Leibler divergence. We derive policy gradient s where the change in policy is limited to a small Wasserstein distance (or trus t region). This is done in the discrete and continuous multi-armed bandit settin gs with entropy regularisation. We show that in the small steps limit with respe

ct to the Wasserstein distance W_2, policy dynamics are governed by the heat equation, following the Jordan-Kinderlehrer-Otto result. This means that policies undergo diffusion and advection, concentrating near actions with high reward. This helps elucidate the nature of convergence in the probability matching setup, and provides justification for empirical practices such as Gaussian policy priors and additive gradient noise.
****
Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks
Pratik Chaudhari,Stefano Soatto
Stochastic gradient descent (SGD) is widely believed to perform implicit regularization when used to train deep neural networks, but the precise manner in which this occurs has thus far been elusive. We prove that SGD minimizes an average potential over the posterior distribution of weights along with an entropic regularization term. This potential is however not the original loss function in general. So SGD does perform variational inference, but for a different loss than the one used to compute the gradients. Even more surprisingly, SGD does not even converge in the classical sense: we show that the most likely trajectories of SGD for deep networks do not behave like Brownian motion around critical points. Instead, they resemble closed loops with deterministic components. We prove that such out-of-equilibrium behavior is a consequence of highly non-isotropic gradient noise in SGD; the covariance matrix of mini-batch gradients for deep networks has a rank as small as 1% of its dimension. We provide extensive empirical validation of these claims, proven in the appendix.
****
On the limitations of first order approximation in GAN dynamics
Jerry Li,Aleksander Madry,John Peebles,Ludwig Schmidt
Generative Adversarial Networks (GANs) have been proposed as an approach to learning generative models. While GANs have demonstrated promising performance on multiple vision tasks, their learning dynamics are not yet well understood, neither in theory nor in practice. In particular, the work in this domain has been focused so far only on understanding the properties of the stationary solutions that this dynamics might converge to, and of the behavior of that dynamics in this solutions' immediate neighborhood.

To address this issue, in this work we take a first step towards a principled study of the GAN dynamics itself. To this end, we propose a model that, on one hand, exhibits several of the common problematic convergence behaviors (e.g., vanishing gradient, mode collapse, diverging or oscillatory behavior), but on the other hand, is sufficiently simple to enable rigorous convergence analysis.

This methodology enables us to exhibit an interesting phenomena: a GAN with an optimal discriminator provably converges, while guiding the GAN training using only a first order approximation of the discriminator leads to unstable GAN dynamics and mode collapse. This suggests that such usage of the first order approximation of the discriminator, which is a de-facto standard in all the existing GAN dynamics, might be one of the factors that makes GAN training so challenging in practice. Additionally, our convergence result constitutes the first rigorous analysis of a dynamics of a concrete parametric GAN.
****
Distributed Fine-tuning of Language Models on Private Data
Vadim Popov,Mikhail Kudinov,Irina Piontkovskaya,Petr Vytovtov,Alex Nevidomsky
One of the big challenges in machine learning applications is that training data can be different from the real-world data faced by the algorithm. In language modeling, users' language (e.g. in private messaging) could change in a year and be completely different from what we observe in publicly available data. At the same time, public data can be used for obtaining general knowledge (i.e. general model of English). We study approaches to distributed fine-tuning of a general model on user private data with the additional requirements of maintaining the quality on the general data and minimization of communication costs. We propose a

novel technique that significantly improves prediction quality on users' langua ge compared to a general model and outperforms gradient compression methods in t erms of communication efficiency. The proposed procedure is fast and leads to an almost 70% perplexity reduction and 8.7 percentage point improvement in keystro ke saving rate on informal English texts. Finally, we propose an experimental fr amework for evaluating differential privacy of distributed training of language models and show that our approach has good privacy guarantees.
****

Noisy Networks For Exploration
Meire Fortunato,Mohammad Gheshlaghi Azar,Bilal Piot,Jacob Menick,Matteo Hessel,I an Osband,Alex Graves,Volodymyr Mnih,Remi Munos,Demis Hassabis,Olivier Pietquin, Charles Blundell,Shane Legg
We introduce NoisyNet, a deep reinforcement learning agent with parametric noise added to its weights, and show that the induced stochasticity of the agent's po licy can be used to aid efficient exploration. The parameters of the noise are l earned with gradient descent along with the remaining network weights. NoisyNet is straightforward to implement and adds little computational overhead. We find that replacing the conventional exploration heuristics for A3C, DQN and Dueling agents (entropy reward and epsilon-greedy respectively) with NoisyNet yields su bstantially higher scores for a wide range of Atari games, in some cases advanci ng the agent from sub to super-human performance.
****

Neural-Guided Deductive Search for Real-Time Program Synthesis from Examples
Ashwin Kalyan,Abhishek Mohta,Oleksandr Polozov,Dhruv Batra,Prateek Jain,Sumit Gu lwani
Synthesizing user-intended programs from a small number of input-output exam-
ples is a challenging problem with several important applications like spreadshe et
manipulation, data wrangling and code refactoring. Existing synthesis systems
either completely rely on deductive logic techniques that are extensively hand-
engineered or on purely statistical models that need massive amounts of data, an d in
general fail to provide real-time synthesis on challenging benchmarks. In this w ork,
we propose Neural Guided Deductive Search (NGDS), a hybrid synthesis technique
that combines the best of both symbolic logic techniques and statistical models.
Thus, it produces programs that satisfy the provided specifications by construct ion
and generalize well on unseen examples, similar to data-driven systems. Our
technique effectively utilizes the deductive search framework to reduce the lear ning
problem of the neural component to a simple supervised learning setup. Further,
this allows us to both train on sparingly available real-world data and still le verage
powerful recurrent neural network encoders. We demonstrate the effectiveness
of our method by evaluating on real-world customer scenarios by synthesizing
accurate programs with up to 12× speed-up compared to state-of-the-art systems.
****

Hyperparameter optimization: a spectral approach
Elad Hazan,Adam Klivans,Yang Yuan
We give a simple, fast algorithm for hyperparameter optimization inspired by tec hniques from the analysis of Boolean functions. We focus on the high-dimensiona l regime where the canonical example is training a neural network with a large n umber of hyperparameters. The algorithm --- an iterative application of compress ed sensing techniques for orthogonal polynomials --- requires only uniform sampl ing of the hyperparameters and is thus easily parallelizable.

Experiments for training deep neural networks on Cifar-10 show that compared to state-of-the-art tools (e.g., Hyperband and Spearmint), our algorithm finds sign ificantly improved solutions, in some cases better than what is attainable by ha

nd-tuning. In terms of overall running time (i.e., time required to sample vari
ous settings of hyperparameters plus additional computation time), we are at lea
st an order of magnitude faster than Hyperband and Bayesian Optimization. We al
so outperform Random Search $8\times$.

Our method is inspired by provably-efficient algorithms for learning decision tr
ees using the discrete Fourier transform. We obtain improved sample-complexty b
ounds for learning decision trees while matching state-of-the-art bounds on runn
ing time (polynomial and quasipolynomial, respectively).
****
Large Batch Training of Convolutional Networks with Layer-wise Adaptive Rate Sca
ling
Boris Ginsburg,Igor Gitman,Yang You
A common way to speed up training of large convolutional networks is to add com
putational units. Training is then performed using data-parallel synchronous Sto
chastic Gradient Descent (SGD) with a mini-batch divided between computational u
nits. With an increase in the number of nodes, the batch size grows. However, t
raining with a large batch often results in lower model accuracy. We argue that
 the current recipe for large batch training (linear learning rate scaling with
warm-up) is not general enough and training may diverge. To overcome these optim
ization difficulties, we propose a new training algorithm based on Layer-wise Ad
aptive Rate Scaling (LARS). Using LARS, we scaled AlexNet and ResNet-50 to a ba
tch size of 16K.
****
Deep Learning with Logged Bandit Feedback
Thorsten Joachims,Adith Swaminathan,Maarten de Rijke
We propose a new output layer for deep neural networks that permits the use of l
ogged contextual bandit feedback for training. Such contextual bandit feedback c
an be available in huge quantities (e.g., logs of search engines, recommender sy
stems) at little cost, opening up a path for training deep networks on orders of
 magnitude more data. To this effect, we propose a Counterfactual Risk Minimizat
ion (CRM) approach for training deep networks using an equivariant empirical ris
k estimator with variance regularization, BanditNet, and show how the resulting
objective can be decomposed in a way that allows Stochastic Gradient Descent (SG
D) training. We empirically demonstrate the effectiveness of the method by showi
ng how deep networks -- ResNets in particular -- can be trained for object recog
nition without conventionally labeled images.
****
Regularization for Deep Learning: A Taxonomy
Jan Kuka■ka,Vladimir Golkov,Daniel Cremers
Regularization is one of the crucial ingredients of deep learning, yet the term
regularization has various definitions, and regularization methods are often stu
died separately from each other. In our work we present a novel, systematic, uni
fying taxonomy to categorize existing methods. We distinguish methods that affec
t data, network architectures, error terms, regularization terms, and optimizati
on procedures. We identify the atomic building blocks of existing methods, and d
ecouple the assumptions they enforce from the mathematical tools they rely on. W
e do not provide all details about the listed methods; instead, we present an ov
erview of how the methods can be sorted into meaningful categories and sub-categ
ories. This helps revealing links and fundamental similarities between them. Fin
ally, we include practical recommendations both for users and for developers of
new regularization methods.
****
Few-shot Autoregressive Density Estimation: Towards Learning to Learn Distributi
ons
Scott Reed,Yutian Chen,Thomas Paine,Aäron van den Oord,S. M. Ali Eslami,Danilo R
ezende,Oriol Vinyals,Nando de Freitas
Deep autoregressive models have shown state-of-the-art performance in density es
timation for natural images on large-scale datasets such as ImageNet. However,
such models require many thousands of gradient-based weight updates and unique i

mage examples for training. Ideally, the models would rapidly learn visual concepts from only a handful of examples, similar to the manner in which humans learns across many vision tasks. In this paper, we show how 1) neural attention and 2) meta learning techniques can be used in combination with autoregressive models to enable effective few-shot density estimation. Our proposed modifications to PixelCNN result in state-of-the art few-shot density estimation on the Omniglot dataset. Furthermore, we visualize the learned attention policy and find that it learns intuitive algorithms for simple tasks such as image mirroring on ImageNet and handwriting on Omniglot without supervision. Finally, we extend the model to natural images and demonstrate few-shot image generation on the Stanford Online Products dataset.
****
TreeQN and ATreeC: Differentiable Tree-Structured Models for Deep Reinforcement Learning
Gregory Farquhar,Tim Rocktäschel,Maximilian Igl,Shimon Whiteson
Combining deep model-free reinforcement learning with on-line planning is a promising approach to building on the successes of deep RL. On-line planning with look-ahead trees has proven successful in environments where transition models are known a priori. However, in complex environments where transition models need to be learned from data, the deficiencies of learned models have limited their utility for planning. To address these challenges, we propose TreeQN, a differentiable, recursive, tree-structured model that serves as a drop-in replacement for any value function network in deep RL with discrete actions. TreeQN dynamically constructs a tree by recursively applying a transition model in a learned abstract state space and then aggregating predicted rewards and state-values using a tree backup to estimate Q-values. We also propose ATreeC, an actor-critic variant that augments TreeQN with a softmax layer to form a stochastic policy network. Both approaches are trained end-to-end, such that the learned model is optimised for its actual use in the tree. We show that TreeQN and ATreeC outperform n-step DQN and A2C on a box-pushing task, as well as n-step DQN and value prediction networks (Oh et al., 2017) on multiple Atari games. Furthermore, we present ablation studies that demonstrate the effect of different auxiliary losses on learning transition models.
****
LEARNING TO SHARE: SIMULTANEOUS PARAMETER TYING AND SPARSIFICATION IN DEEP LEARNING
Dejiao Zhang,Haozhu Wang,Mario Figueiredo,Laura Balzano
Deep neural networks (DNNs) usually contain millions, maybe billions, of parameters/weights, making both storage and computation very expensive. This has motivated a large body of work to reduce the complexity of the neural network by using sparsity-inducing regularizers. Another well-known approach for controlling the complexity of DNNs is parameter sharing/tying, where certain sets of weights are forced to share a common value. Some forms of weight sharing are hard-wired to express certain in- variances, with a notable example being the shift-invariance of convolutional layers. However, there may be other groups of weights that may be tied together during the learning process, thus further re- ducing the complexity of the network. In this paper, we adopt a recently proposed sparsity-inducing regularizer, named GrOWL (group ordered weighted l1), which encourages sparsity and, simulta- neously, learns which groups of parameters should share a common value. GrOWL has been proven effective in linear regression, being able to identify and cope with strongly correlated covariates. Unlike standard sparsity-inducing regularizers (e.g., l1 a.k.a. Lasso), GrOWL not only eliminates unimportant neurons by setting all the corresponding weights to zero, but also explicitly identifies strongly correlated neurons by tying the corresponding weights to a common value. This ability of GrOWL motivates the following two-stage procedure: (i) use GrOWL regularization in the training process to simultaneously identify significant neurons and groups of parameter that should be tied together; (ii) retrain the network, enforcing the structure that was unveiled in the previous phase, i.e., keeping only the significant neurons and enforcing the learned tying structure. We evaluate the proposed approach on several benchmark datasets, s

howing that it can dramatically compress the network with slight or even no loss on generalization performance.

****

Realtime query completion via deep language models
Po-Wei Wang,J. Zico Kolter,Vijai Mohan,Inderjit S. Dhillon
Search engine users nowadays heavily depend on query completion and correction to shape their queries. Typically, the completion is done by database lookup which does not understand the context and cannot generalize to prefixes not in the database. In the paper, we propose to use unsupervised deep language models to complete and correct the queries given an arbitrary prefix. We show how to address two main challenges that renders this method practical for large-scale deployment: 1) we propose a method for integrating error correction into the language model completion via a edit-distance potential and a variant of beam search that can exploit these potential functions; and 2) we show how to efficiently perform CPU-based computation to complete the queries, with error correction, in real time (generating top 10 completions within 16 ms). Experiments show that the method substantially increases hit rate over standard approaches, and is capable of handling tail queries.

****

UNSUPERVISED METRIC LEARNING VIA NONLINEAR FEATURE SPACE TRANSFORMATIONS
Pin Zhang,Bibo Shi,JundongLiu
In this paper, we propose a nonlinear unsupervised metric learning framework to boost of the performance of clustering algorithms. Under our framework, nonlinear distance metric learning and manifold embedding are integrated and conducted simultaneously to increase the natural separations among data samples. The metric learning component is implemented through feature space transformations, regulated by a nonlinear deformable model called Coherent Point Drifting (CPD). Driven by CPD, data points can get to a higher level of linear separability, which is subsequently picked up by the manifold embedding component to generate well-separable sample projections for clustering. Experimental results on synthetic and benchmark datasets show the effectiveness of our proposed approach over the state-of-the-art solutions in unsupervised metric learning.

****

Adversarial Examples for Natural Language Classification Problems
Volodymyr Kuleshov,Shantanu Thakoor,Tingfung Lau,Stefano Ermon
Modern machine learning algorithms are often susceptible to adversarial examples — maliciously crafted inputs that are undetectable by humans but that fool the algorithm into producing undesirable behavior. In this work, we show that adversarial examples exist in natural language classification: we formalize the notion of an adversarial example in this setting and describe algorithms that construct such examples. Adversarial perturbations can be crafted for a wide range of tasks — including spam filtering, fake news detection, and sentiment analysis — and affect different models — convolutional and recurrent neural networks as well as linear classifiers to a lesser degree. Constructing an adversarial example involves replacing 10-30% of words in a sentence with synonyms that don’t change its meaning. Up to 90% of input examples admit adversarial perturbations; furthermore, these perturbations retain a degree of transferability across models. Our findings demonstrate the existence of vulnerabilities in machine learning systems and hint at limitations in our understanding of classification algorithms.
****

Sparse Regularized Deep Neural Networks For Efficient Embedded Learning
Jia Bi
Deep learning is becoming more widespread in its application due to its power in solving complex classification problems. However, deep learning models often require large memory and energy consumption, which may prevent them from being deployed effectively on embedded platforms, limiting their applications. This work addresses the problem by proposing methods {\em Weight Reduction Quantisation} f

or compressing the memory footprint of the models, including reducing the number
 of weights and the number of bits to store each weight. Beside, applying with s
parsity-inducing regularization, our work focuses on speeding up stochastic vari
ance reduced gradients (SVRG) optimization on non-convex problem. Our method tha
t mini-batch SVRG with ℓ1 regularization on non-convex problem has faster a
nd smoother convergence rates than SGD by using adaptive learning rates. Experim
ental evaluation of our approach uses MNIST and CIFAR-10 datasets on LeNet-300-1
00 and LeNet-5 models, showing our approach can reduce the memory requirements b
oth in the convolutional and fully connected layers by up to 60$\times$ without
affecting their test accuracy.
****
Inducing Grammars with and for Neural Machine Translation
Ke Tran,Yonatan Bisk
Previous work has demonstrated the benefits of incorporating additional linguist
ic annotations such as syntactic trees into neural machine translation. However
the cost of obtaining those syntactic annotations is expensive for many language
s and the quality of unsupervised learning linguistic structures is too poor to
be helpful. In this work, we aim to improve neural machine translation via sourc
e side dependency syntax but without explicit annotation. We propose a set of mo
dels that learn to induce dependency trees on the source side and learn to use t
hat information on the target side. Importantly, we also show that our dependenc
y trees capture important syntactic features of language and improve translation
 quality on two language pairs En-De and En-Ru.
****
A New Method of Region Embedding for Text Classification
chao qiao,bo huang,guocheng niu,daren li,daxiang dong,wei he,dianhai yu,hua wu
To represent a text as a bag of properly identified “phrases” and use the repres
entation for processing the text is proved to be useful. The key question here i
s how to identify the phrases and represent them. The traditional method of util
izing n-grams can be regarded as an approximation of the approach. Such a method
 can suffer from data sparsity, however, particularly when the length of n-gram
is large. In this paper, we propose a new method of learning and utilizing task-
specific distributed representations of n-grams, referred to as “region embeddin
gs”. Without loss of generality we address text classification. We specifically
propose two models for region embeddings. In our models, the representation of a
 word has two parts, the embedding of the word itself, and a weighting matrix to
 interact with the local context, referred to as local context unit. The region
embeddings are learned and used in the classification task, as parameters of the
 neural network classifier. Experimental results show that our proposed method o
utperforms existing methods in text classification on several benchmark datasets
. The results also indicate that our method can indeed capture the salient phras
al expressions in the texts.
****
Byte-Level Recursive Convolutional Auto-Encoder for Text
Xiang Zhang,Yann LeCun
This article proposes to auto-encode text at byte-level using convolutional netw
orks with a recursive architecture. The motivation is to explore whether it is p
ossible to have scalable and homogeneous text generation at byte-level in a non-
sequential fashion through the simple task of auto-encoding. We show that non-se
quential text generation from a fixed-length representation is not only possible
, but also achieved much better auto-encoding results than recurrent networks. T
he proposed model is a multi-stage deep convolutional encoder-decoder framework
using residual connections, containing up to 160 parameterized layers. Each enco
der or decoder contains a shared group of modules that consists of either poolin
g or upsampling layers, making the network recursive in terms of abstraction lev
els in representation. Results for 6 large-scale paragraph datasets are reported
, in 3 languages including Arabic, Chinese and English. Analyses are conducted t
o study several properties of the proposed model.
****
Enhancing Batch Normalized Convolutional Networks using Displaced Rectifier Line

ar Units: A Systematic Comparative Study
David Macêdo,Cleber Zanchettin,Adriano L. I. Oliveira,Teresa Ludermir
In this paper, we turn our attention to the interworking between the activation functions and the batch normalization, which is a virtually mandatory technique to train deep networks currently. We propose the activation function Displaced Rectifier Linear Unit (DReLU) by conjecturing that extending the identity function of ReLU to the third quadrant enhances compatibility with batch normalization. Moreover, we used statistical tests to compare the impact of using distinct activation functions (ReLU, LReLU, PReLU, ELU, and DReLU) on the learning speed and test accuracy performance of standardized VGG and Residual Networks state-of-the-art models. These convolutional neural networks were trained on CIFAR-100 and CIFAR-10, the most commonly used deep learning computer vision datasets. The results showed DReLU speeded up learning in all models and datasets. Besides, statistical significant performance assessments (p<0.05) showed DReLU enhanced the test accuracy presented by ReLU in all scenarios. Furthermore, DReLU showed better test accuracy than any other tested activation function in all experiments with one exception, in which case it presented the second best performance. Therefore, this work demonstrates that it is possible to increase performance replacing ReLU by an enhanced activation function.
****
Deep Voice 3: Scaling Text-to-Speech with Convolutional Sequence Learning
Wei Ping,Kainan Peng,Andrew Gibiansky,Sercan O. Arik,Ajay Kannan,Sharan Narang,Jonathan Raiman,John Miller
We present Deep Voice 3, a fully-convolutional attention-based neural text-to-speech (TTS) system. Deep Voice 3 matches state-of-the-art neural speech synthesis systems in naturalness while training an order of magnitude faster. We scale Deep Voice 3 to dataset sizes unprecedented for TTS, training on more than eight hundred hours of audio from over two thousand speakers. In addition, we identify common error modes of attention-based speech synthesis networks, demonstrate how to mitigate them, and compare several different waveform synthesis methods. We also describe how to scale inference to ten million queries per day on a single GPU server.
****
Combination of Supervised and Reinforcement Learning For Vision-Based Autonomous Control
Dmitry Kangin,Nicolas Pugeault
 Reinforcement learning methods have recently achieved impressive results on a wide range of control problems. However, especially with complex inputs, they still require an extensive amount of training data in order to converge to a meaningful solution. This limitation largely prohibits their usage for complex input spaces such as video signals, and it is still impossible to use it for a number of complex problems in a real world environments, including many of those for video based control. Supervised learning, on the contrary, is capable of learning on a relatively small number of samples, however it does not take into account reward-based control policies and is not capable to provide independent control policies. In this article we propose a model-free control method, which uses a combination of reinforcement and supervised learning for autonomous control and paves the way towards policy based control in real world environments. We use SpeedDreams/TORCS video game to demonstrate that our approach requires much less samples (hundreds of thousands against millions or tens of millions) comparing to the state-of-the-art reinforcement learning techniques on similar data, and at the same time overcomes both supervised and reinforcement learning approaches in terms of quality. Additionally, we demonstrate the applicability of the method to MuJoCo control problems.
****
HexaConv
Emiel Hoogeboom,Jorn W.T. Peters,Taco S. Cohen,Max Welling
The effectiveness of Convolutional Neural Networks stems in large part from their ability to exploit the translation invariance that is inherent in many learning problems. Recently, it was shown that CNNs can exploit other invariances, such

as rotation invariance, by using group convolutions instead of planar convoluti ons. However, for reasons of performance and ease of implementation, it has been necessary to limit the group convolution to transformations that can be applied to the filters without interpolation. Thus, for images with square pixels, only integer translations, rotations by multiples of 90 degrees, and reflections are admissible.

Whereas the square tiling provides a 4-fold rotational symmetry, a hexagonal til ing of the plane has a 6-fold rotational symmetry. In this paper we show how one can efficiently implement planar convolution and group convolution over hexagon al lattices, by re-using existing highly optimized convolution routines. We find that, due to the reduced anisotropy of hexagonal filters, planar HexaConv provi des better accuracy than planar convolution with square filters, given a fixed p arameter budget. Furthermore, we find that the increased degree of symmetry of t he hexagonal grid increases the effectiveness of group convolutions, by allowing for more parameter sharing. We show that our method significantly outperforms c onventional CNNs on the AID aerial scene classification dataset, even outperform ing ImageNet pre-trained models.
****

Coulomb GANs: Provably Optimal Nash Equilibria via Potential Fields

Thomas Unterthiner,Bernhard Nessler,Calvin Seward,Günter Klambauer,Martin Heusel ,Hubert Ramsauer,Sepp Hochreiter

Generative adversarial networks (GANs) evolved into one of the most successful u nsupervised techniques for generating realistic images. Even though it has recen tly been shown that GAN training converges, GAN models often end up in local Nas h equilibria that are associated with mode collapse or otherwise fail to model t he target distribution. We introduce Coulomb GANs, which pose the GAN learning p roblem as a potential field, where generated samples are attracted to training s et samples but repel each other. The discriminator learns a potential field whil e the generator decreases the energy by moving its samples along the vector (for ce) field determined by the gradient of the potential field. Through decreasing the energy, the GAN model learns to generate samples according to the whole targ et distribution and does not only cover some of its modes. We prove that Coulomb GANs possess only one Nash equilibrium which is optimal in the sense that the m odel distribution equals the target distribution. We show the efficacy of Coulom b GANs on LSUN bedrooms, CelebA faces, CIFAR-10 and the Google Billion Word text generation.
****

Acquiring Target Stacking Skills by Goal-Parameterized Deep Reinforcement Learni ng

Wenbin Li,Jeannette Bohg,Mario Fritz

Understanding physical phenomena is a key component of human intelligence and en ables physical interaction with previously unseen environments. In this paper, w e study how an artificial agent can autonomously acquire this intuition through interaction with the environment. We created a synthetic block stacking environm ent with physics simulation in which the agent can learn a policy end-to-end thr ough trial and error. Thereby, we bypass to explicitly model physical knowledge within the policy. We are specifically interested in tasks that require the agen t to reach a given goal state that may be different for every new trial. To this end, we propose a deep reinforcement learning framework that learns policies wh ich are parametrized by a goal. We validated the model on a toy example navigati ng in a grid world with different target positions and in a block stacking task with different target structures of the final tower. In contrast to prior work, our policies show better generalization across different goals.
****

Counterfactual Image Networks

Deniz Oktay,Carl Vondrick,Antonio Torralba

We capitalize on the natural compositional structure of images in order to learn object segmentation with weakly labeled images. The intuition behind our approa ch is that removing objects from images will yield natural images, however remov

ing random patches will yield unnatural images. We leverage this signal to devel
op a generative model that decomposes an image into layers, and when all layers
are combined, it reconstructs the input image. However, when a layer is removed,
 the model learns to produce a different image that still looks natural to an ad
versary, which is possible by removing objects. Experiments and visualizations s
uggest that this model automatically learns object segmentation on images labele
d only by scene better than baselines.
****
Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a
 Stochastic Actor
Tuomas Haarnoja,Aurick Zhou,Pieter Abbeel,Sergey Levine
Model-free deep reinforcement learning (RL) algorithms have been demonstrated on
 a range of challenging decision making and control tasks. However, these method
s typically suffer from two major challenges: very high sample complexity and br
ittle convergence properties, which necessitate meticulous hyperparameter tuning
. Both of these challenges severely limit the applicability of such methods to c
omplex, real-world domains. In this paper, we propose soft actor-critic, an off-
policy actor-critic deep RL algorithm based on the maximum entropy reinforcement
 learning framework. In this framework, the actor aims to maximize expected rewa
rd while also maximizing entropy - that is, succeed at the task while acting as
randomly as possible. Prior deep RL methods based on this framework have been fo
rmulated as either off-policy Q-learning, or on-policy policy gradient methods.
By combining off-policy updates with a stable stochastic actor-critic formulatio
n, our method achieves state-of-the-art performance on a range of continuous con
trol benchmark tasks, outperforming prior on-policy and off-policy methods. Furt
hermore, we demonstrate that, in contrast to other off-policy algorithms, our ap
proach is very stable, achieving very similar performance across different rando
m seeds.
****
Neural Sketch Learning for Conditional Program Generation
Vijayaraghavan Murali,Letao Qi,Swarat Chaudhuri,Chris Jermaine
We study the problem of generating source code in a strongly typed,
Java-like programming language, given a label (for example a set of
API calls or types) carrying a small amount of information about the
code that is desired. The generated programs are expected to respect a
`"realistic" relationship between programs and labels, as exemplified
by a corpus of labeled programs available during training.

Two challenges in such *conditional program generation* are that
the generated programs must satisfy a rich set of syntactic and
semantic constraints, and that source code contains many low-level
features that impede learning. We address these problems by training
a neural generator not on code but on *program sketches*, or
models of program syntax that abstract out names and operations that
do not generalize across programs. During generation, we infer a
posterior distribution over sketches, then concretize samples from
this distribution into type-safe programs using combinatorial
techniques. We implement our ideas in a system for generating
API-heavy Java code, and show that it can often predict the entire
body of a method given just a few API calls or data types that appear
in the method.
****
ShakeDrop regularization
Yoshihiro Yamada,Masakazu Iwamura,Koichi Kise
This paper proposes a powerful regularization method named \textit{ShakeDrop reg
ularization}.
ShakeDrop is inspired by Shake-Shake regularization that decreases error rates b
y disturbing learning.
While Shake-Shake can be applied to only ResNeXt which has multiple branches, Sh
akeDrop can be applied to not only ResNeXt but also ResNet, Wide ResNet and Pyra

midNet in a memory efficient way.
Important and interesting feature of ShakeDrop is that it strongly disturbs learning by multiplying even a negative factor to the output of a convolutional layer in the forward training pass.
The effectiveness of ShakeDrop is confirmed by experiments on CIFAR-10/100 and Tiny ImageNet datasets.

****

Predict Responsibly: Increasing Fairness by Learning to Defer
David Madras,Toniann Pitassi,Richard Zemel
When machine learning models are used for high-stakes decisions, they should predict accurately, fairly, and responsibly. To fulfill these three requirements, a model must be able to output a reject option (i.e. say "``I Don't Know") when it is not qualified to make a prediction. In this work, we propose learning to defer, a method by which a model can defer judgment to a downstream decision-maker such as a human user. We show that learning to defer generalizes the rejection learning framework in two ways: by considering the effect of other agents in the decision-making process, and by allowing for optimization of complex objectives. We propose a learning algorithm which accounts for potential biases held by decision-makerslater in a pipeline. Experiments on real-world datasets demonstrate that learning
to defer can make a model not only more accurate but also less biased. Even when operated by highly biased users, we show that
deferring models can still greatly improve the fairness of the entire pipeline.
****

Learn to Pay Attention
Saumya Jetley,Nicholas A. Lord,Namhoon Lee,Philip H. S. Torr
We propose an end-to-end-trainable attention module for convolutional neural network (CNN) architectures built for image classification. The module takes as input the 2D feature vector maps which form the intermediate representations of the input image at different stages in the CNN pipeline, and outputs a 2D matrix of scores for each map. Standard CNN architectures are modified through the incorporation of this module, and trained under the constraint that a convex combination of the intermediate 2D feature vectors, as parametrised by the score matrices, must alone be used for classification. Incentivised to amplify the relevant and suppress the irrelevant or misleading, the scores thus assume the role of attention values. Our experimental observations provide clear evidence to this effect: the learned attention maps neatly highlight the regions of interest while suppressing background clutter. Consequently, the proposed function is able to bootstrap standard CNN architectures for the task of image classification, demonstrating superior generalisation over 6 unseen benchmark datasets. When binarised, our attention maps outperform other CNN-based attention maps, traditional saliency maps, and top object proposals for weakly supervised segmentation as demonstrated on the Object Discovery dataset. We also demonstrate improved robustness against the fast gradient sign method of adversarial attack.
****

Estimation of cross-lingual news similarities using text-mining methods
Zhouhao Wang,Enda Liu,Hiroki Sakaji,Tomoki Ito,Kiyoshi Izumi,Kota Tsubouchi,Tatsuo Yamashita
Every second, innumerable text data, including all kinds news, reports, messages, reviews, comments, and twits have been generated on the Internet, which is written not only in English but also in other languages such as Chinese, Japanese, French and so on. Not only SNS sites but also worldwide news agency such as Thomson Reuters News provide news reported in more than 20 languages, reflecting the significance of the multilingual information.
In this research, by taking advantage of multi-lingual text resources provided by the Thomson Reuters News, we developed a bidirectional LSTM based method to calculate cross-lingual semantic text similarity for long text and short text respectively. Thus, users could understand the situation comprehensively, by investigating similar and related cross-lingual articles, when there an important news

comes in.
****
SMASH: One-Shot Model Architecture Search through HyperNetworks
Andrew Brock,Theo Lim,J.M. Ritchie,Nick Weston
Designing architectures for deep neural networks requires expert knowledge and substantial computation time. We propose a technique to accelerate architecture selection by learning an auxiliary HyperNet that generates the weights of a main model conditioned on that model's architecture. By comparing the relative validation performance of networks with HyperNet-generated weights, we can effectively search over a wide range of architectures at the cost of a single training run. To facilitate this search, we develop a flexible mechanism based on memory read-writes that allows us to define a wide range of network connectivity patterns, with ResNet, DenseNet, and FractalNet blocks as special cases. We validate our method (SMASH) on CIFAR-10 and CIFAR-100, STL-10, ModelNet10, and Imagenet32x32, achieving competitive performance with similarly-sized hand-designed networks.
****
On the regularization of Wasserstein GANs
Henning Petzka,Asja Fischer,Denis Lukovnikov
Since their invention, generative adversarial networks (GANs) have become a popular approach for learning to model a distribution of real (unlabeled) data. Convergence problems during training are overcome by Wasserstein GANs which minimize the distance between the model and the empirical distribution in terms of a different metric, but thereby introduce a Lipschitz constraint into the optimization problem. A simple way to enforce the Lipschitz constraint on the class of functions, which can be modeled by the neural network, is weight clipping. Augmenting the loss by a regularization term that penalizes the deviation of the gradient norm of the critic (as a function of the network's input) from one, was proposed as an alternative that improves training. We present theoretical arguments why using a weaker regularization term enforcing the Lipschitz constraint is preferable. These arguments are supported by experimental results on several data sets.
****
INTERPRETATION OF NEURAL NETWORK IS FRAGILE
Amirata Ghorbani,Abubakar Abid,James Zou
In order for machine learning to be deployed and trusted in many applications, it is crucial to be able to reliably explain why the machine learning algorithm makes certain predictions. For example, if an algorithm classifies a given pathology image to be a malignant tumor, then the doctor may need to know which parts of the image led the algorithm to this classification. How to interpret black-box predictors is thus an important and active area of research. A fundamental question is: how much can we trust the interpretation itself? In this paper, we show that interpretation of deep learning predictions is extremely fragile in the following sense: two perceptively indistinguishable inputs with the same predicted label can be assigned very different}interpretations. We systematically characterize the fragility of the interpretations generated by several widely-used feature-importance interpretation methods (saliency maps, integrated gradient, and DeepLIFT) on ImageNet and CIFAR-10. Our experiments show that even small random perturbation can change the feature importance and new systematic perturbations can lead to dramatically different interpretations without changing the label. We extend these results to show that interpretations based on exemplars (e.g. influence functions) are similarly fragile. Our analysis of the geometry of the Hessian matrix gives insight on why fragility could be a fundamental challenge to the current interpretation approaches.
****
Interactive Grounded Language Acquisition and Generalization in a 2D World
Haonan Yu,Haichao Zhang,Wei Xu
We build a virtual agent for learning language in a 2D maze-like world. The agent sees images of the surrounding environment, listens to a virtual teacher, and takes actions to receive rewards. It interactively learns the teacher’s language from scratch based on two language use cases: sentence-directed navigation and

question answering. It learns simultaneously the visual representations of the world, the language, and the action control. By disentangling language grounding from other computational routines and sharing a concept detection function between language grounding and prediction, the agent reliably interpolates and extrapolates to interpret sentences that contain new word combinations or new words missing from training sentences. The new words are transferred from the answers of language prediction. Such a language ability is trained and evaluated on a population of over 1.6 million distinct sentences consisting of 119 object words, 8 color words, 9 spatial-relation words, and 50 grammatical words. The proposed model significantly outperforms five comparison methods for interpreting zero-shot sentences. In addition, we demonstrate human-interpretable intermediate outputs of the model in the appendix.

****

A Semantic Loss Function for Deep Learning with Symbolic Knowledge

Jingyi Xu,Zilu Zhang,Tal Friedman,Yitao Liang,Guy Van den Broeck

This paper develops a novel methodology for using symbolic knowledge in deep learning. From first principles, we derive a semantic loss function that bridges between neural output vectors and logical constraints. This loss function captures how close the neural network is to satisfying the constraints on its output. An experimental evaluation shows that our semantic loss function effectively guides the learner to achieve (near-)state-of-the-art results on semi-supervised multi-class classification. Moreover, it significantly increases the ability of the neural network to predict structured objects, such as rankings and shortest paths. These discrete concepts are tremendously difficult to learn, and benefit from a tight integration of deep learning and symbolic reasoning methods.

****

Interpretable Classification via Supervised Variational Autoencoders and Differentiable Decision Trees

Eleanor Quint,Garrett Wirka,Jacob Williams,Stephen Scott,N.V. Vinodchandran

As deep learning-based classifiers are increasingly adopted in real-world applications, the importance of understanding how a particular label is chosen grows. Single decision trees are an example of a simple, interpretable classifier, but are unsuitable for use with complex, high-dimensional data. On the other hand, the variational autoencoder (VAE) is designed to learn a factored, low-dimensional representation of data, but typically encodes high-likelihood data in an intrinsically non-separable way. We introduce the differentiable decision tree (DDT) as a modular component of deep networks and a simple, differentiable loss function that allows for end-to-end optimization of a deep network to compress high-dimensional data for classification by a single decision tree. We also explore the power of labeled data in a supervised VAE (SVAE) with a Gaussian mixture prior, which leverages label information to produce a high-quality generative model with improved bounds on log-likelihood. We combine the SVAE with the DDT to get our classifier+VAE (C+VAE), which is competitive in both classification error and log-likelihood, despite optimizing both simultaneously and using a very simple encoder/decoder architecture.

****

Generating Differentially Private Datasets Using GANs

Aleksei Triastcyn,Boi Faltings

In this paper, we present a technique for generating artificial datasets that retain statistical properties of the real data while providing differential privacy guarantees with respect to this data. We include a Gaussian noise layer in the discriminator of a generative adversarial network to make the output and the gradients differentially private with respect to the training data, and then use the generator component to synthesise privacy-preserving artificial dataset. Our experiments show that under a reasonably small privacy budget we are able to generate data of high quality and successfully train machine learning models on this artificial data.

****

Code Synthesis with Priority Queue Training

Daniel A. Abolafia,Quoc V. Le,Mohammad Norouzi

We consider the task of program synthesis in the presence of a reward function o
ver the output of programs, where the goal is to find programs with maximal rewa
rds. We introduce a novel iterative optimization scheme, where we train an RNN o
n a dataset of K best programs from a priority queue of the generated programs s
o far. Then, we synthesize new programs and add them to the priority queue by sa
mpling from the RNN. We benchmark our algorithm called priority queue training (
PQT) against genetic algorithm and reinforcement learning baselines on a simple
but expressive Turing complete programming language called BF. Our experimental
results show that our deceptively simple PQT algorithm significantly outperforms
 the baselines. By adding a program length penalty to the reward function, we ar
e able to synthesize short, human readable programs.
****
On the Information Bottleneck Theory of Deep Learning
Andrew Michael Saxe,Yamini Bansal,Joel Dapello,Madhu Advani,Artemy Kolchinsky,Br
endan Daniel Tracey,David Daniel Cox
The practical successes of deep neural networks have not been matched by theoret
ical progress that satisfyingly explains their behavior. In this work, we study
the information bottleneck (IB) theory of deep learning, which makes three speci
fic claims: first, that deep networks undergo two distinct phases consisting of
an initial fitting phase and a subsequent compression phase; second, that the co
mpression phase is causally related to the excellent generalization performance
of deep networks; and third, that the compression phase occurs due to the diffus
ion-like behavior of stochastic gradient descent. Here we show that none of thes
e claims hold true in the general case. Through a combination of analytical resu
lts and simulation, we demonstrate that the information plane trajectory is pred
ominantly a function of the neural nonlinearity employed: double-sided saturatin
g nonlinearities like tanh yield a compression phase as neural activations enter
 the saturation regime, but linear activation functions and single-sided saturat
ing nonlinearities like the widely used ReLU in fact do not. Moreover, we find t
hat there is no evident causal connection between compression and generalization
: networks that do not compress are still capable of generalization, and vice ve
rsa. Next, we show that the compression phase, when it exists, does not arise fr
om stochasticity in training by demonstrating that we can replicate the IB findi
ngs using full batch gradient descent rather than stochastic gradient descent. F
inally, we show that when an input domain consists of a subset of task-relevant
and task-irrelevant information, hidden representations do compress the task-irr
elevant information, although the overall information about the input may monoto
nically increase with training time, and that this compression happens concurren
tly with the fitting process rather than during a subsequent compression period.
****
Global Convergence of Policy Gradient Methods for Linearized Control Problems
Maryam Fazel,Rong Ge,Sham M. Kakade,Mehran Mesbahi
Direct policy gradient methods for reinforcement learning and continuous control
 problems are a popular
approach for a variety of reasons:
1) they are easy to implement without explicit knowledge of the underlying model
;
2) they are an "end-to-end" approach, directly optimizing the performance metric
 of interest;
3) they inherently allow for richly parameterized policies.
A notable drawback is that even in the most basic continuous control problem (th
at of linear quadratic regulators), these methods must solve a non-convex optimi
zation problem, where little is understood about their efficiency from both comp
utational and statistical perspectives. In contrast, system identification and m
odel based planning in optimal control theory have a much more solid theoretical
 footing, where much is known with regards to their computational and statistica
l properties. This work bridges this gap showing that (model free) policy gradi
ent methods globally converge to the optimal solution and are efficient (polynom
ially so in relevant problem dependent quantities) with regards to their sample
and computational complexities.

****
Causal Generative Neural Networks
Olivier Goudet,Diviyan Kalainathan,David Lopez-Paz,Philippe Caillou,Isabelle Guyon,Michèle Sebag
We introduce CGNN, a framework to learn functional causal models as generative neural networks. These networks are trained using backpropagation to minimize the maximum mean discrepancy to the observed data. Unlike previous approaches, CGNN leverages both conditional independences and distributional asymmetries to seamlessly discover bivariate and multivariate
causal structures, with or without hidden variables. CGNN does not only estimate the causal structure, but a full and differentiable generative model of the data. Throughout an extensive variety of experiments, we illustrate the competitive esults of CGNN w.r.t state-of-the-art alternatives in observational causal discovery on both simulated and real data, in the tasks of cause-effect inference, v-structure identification, and multivariate causal discovery.
****
Adversarial Dropout Regularization
Kuniaki Saito,Yoshitaka Ushiku,Tatsuya Harada,Kate Saenko
We present a domain adaptation method for transferring neural representations from label-rich source domains to unlabeled target domains. Recent adversarial methods proposed for this task learn to align features across domains by ``fooling'' a special domain classifier network. However, a drawback of this approach is that the domain classifier simply labels the generated features as in-domain or not, without considering the boundaries between classes. This means that ambiguous target features can be generated near class boundaries, reducing target classification accuracy. We propose a novel approach, Adversarial Dropout Regularization (ADR), which encourages the generator to output more discriminative features for the target domain. Our key idea is to replace the traditional domain critic with a critic that detects non-discriminative features by using dropout on the classifier network. The generator then learns to avoid these areas of the feature space and thus creates better features. We apply our ADR approach to the problem of unsupervised domain adaptation for image classification and semantic segmentation tasks, and demonstrate significant improvements over the state of the art.
****
SIC-GAN: A Self-Improving Collaborative GAN for Decoding Sketch RNNs
Chi-Chun Chuang,Zheng-Xin Weng,Shan-Hung Wu
Variational RNNs are proposed to output “creative” sequences. Ideally, a collection of sequences produced by a variational RNN should be of both high quality and high variety. However, existing decoders for variational RNNs suffer from a trade-off between quality and variety. In this paper, we seek to learn a variational RNN that decodes high-quality and high-variety sequences. We propose the Self-Improving Collaborative GAN (SIC-GAN), where there are two generators (variational RNNs) collaborating with each other to output a sequence and aiming to trick the discriminator into believing the sequence is of good quality. By deliberately weakening one generator, we can make another stronger in balancing quality and variety. We conduct experiments using the QuickDraw dataset and the results demonstrate the effectiveness of SIC-GAN empirically.
****
Detecting Statistical Interactions from Neural Network Weights
Michael Tsang,Dehua Cheng,Yan Liu
Interpreting neural networks is a crucial and challenging task in machine learning. In this paper, we develop a novel framework for detecting statistical interactions captured by a feedforward multilayer neural network by directly interpreting its learned weights. Depending on the desired interactions, our method can achieve significantly better or similar interaction detection performance compared to the state-of-the-art without searching an exponential solution space of possible interactions. We obtain this accuracy and efficiency by observing that interactions between input features are created by the non-additive effect of nonlinear activation functions, and that interacting paths are encoded in weight matr

ices. We demonstrate the performance of our method and the importance of discove
red interactions via experimental results on both synthetic datasets and real-wo
rld application datasets.
****
Avoiding Catastrophic States with Intrinsic Fear
Zachary C. Lipton,Kamyar Azizzadenesheli,Abhishek Kumar,Lihong Li,Jianfeng Gao,L
i Deng
Many practical reinforcement learning problems contain catastrophic states that
the optimal policy visits infrequently or never. Even on toy problems, deep rein
forcement learners periodically revisit these states, once they are forgotten un
der a new policy. In this paper, we introduce intrinsic fear, a learned reward s
haping that accelerates deep reinforcement learning and guards oscillating polic
ies against periodic catastrophes. Our approach incorporates a second model trai
ned via supervised learning to predict the probability of imminent catastrophe.
This score acts as a penalty on the Q-learning objective. Our theoretical analys
is demonstrates that the perturbed objective yields the same average return unde
r strong assumptions and an ϵ-close average return under weaker assumpt
ions. Our analysis also shows robustness to classification errors. Equipped with
 intrinsic fear, our DQNs solve the toy environments and improve on the Atari ga
mes Seaquest, Asteroids, and Freeway.
****
Siamese Survival Analysis with Competing Risks
Anton Nemchenko,Kartik Ahuja,Mihaela Van Der Schaar
Survival Analysis (time-to-event analysis) in the presence of multiple possible
adverse events, i.e., competing risks, is a challenging, yet very important prob
lem in medicine, finance, manufacturing, etc. Extending classical survival analy
sis to competing risks is not trivial since only one event (e.g. one cause of de
ath) is observed and hence, the incidence of an event of interest is often obscu
red by other related competing events. This leads to the nonidentifiability of t
he event times’ distribution parameters, which makes the problem significantly m
ore challenging. In this work we introduce Siamese Survival Prognosis Network, a
 novel Siamese Deep Neural Network architecture that is able to effectively lear
n from data in the presence of multiple adverse events. The Siamese Survival Net
work is especially crafted to issue pairwise concordant time-dependent risks, in
 which longer event times are assigned lower risks. Furthermore, our architectur
e is able to directly optimize an approximation to the C-discrimination index, r
ather than relying on well-known metrics of cross-entropy etc., and which are no
t able to capture the unique requirements of survival analysis with competing ri
sks. Our results show consistent performance improvements on a number of publicl
y available medical datasets over both statistical and deep learning state-of-th
e-art methods.
****
Reinforcement Learning via Replica Stacking of Quantum Measurements for the Trai
ning of Quantum Boltzmann Machines
Anna Levit,■ Daniel Crawford,Navid Ghadermarzy,Jaspreet S. Oberoi,Ehsan Zahedine
jad,Pooya Ronagh
Recent theoretical and experimental results suggest the possibility of using cur
rent and near-future quantum hardware in challenging sampling tasks. In this pap
er, we introduce free-energy-based reinforcement learning (FERL) as an applicati
on of quantum hardware. We propose a method for processing a quantum annealer’s
measured qubit spin configurations in approximating the free energy of a quantum
 Boltzmann machine (QBM). We then apply this method to perform reinforcement lea
rning on the grid-world problem using the D-Wave 2000Q quantum annealer. The exp
erimental results show that our technique is a promising method for harnessing t
he power of quantum sampling in reinforcement learning tasks.
****
Kernel Graph Convolutional Neural Nets
Giannis Nikolentzos,Polykarpos Meladianos,Antoine J-P Tixier,Konstantinos Skiani
s,Michalis Vazirgiannis
Graph kernels have been successfully applied to many graph classification proble

ms. Typically, a kernel is first designed, and then an SVM classifier is trained
 based on the features defined implicitly by this kernel. This two-stage approac
h decouples data representation from learning, which is suboptimal. On the other
 hand, Convolutional Neural Networks (CNNs) have the capability to learn their o
wn features directly from the raw data during training. Unfortunately, they cann
ot handle irregular data such as graphs. We address this challenge by using grap
h kernels to embed meaningful local neighborhoods of the graphs in a continuous
vector space. A set of filters is then convolved with these patches, pooled, and
 the output is then passed to a feedforward network. With limited parameter tuni
ng, our approach outperforms strong baselines on 7 out of 10 benchmark datasets,
 and reaches comparable performance elsewhere. Code and data are publicly availa
ble.
****
On the Use of Word Embeddings Alone to Represent Natural Language Sequences
Dinghan Shen,Guoyin Wang,Wenlin Wang,Martin Renqiang Min,Qinliang Su,Yizhe Zhang
,Ricardo Henao,Lawrence Carin
To construct representations for natural language sequences, information from tw
o main sources needs to be captured: (i) semantic meaning of individual words, a
nd (ii) their compositionality. These two types of information are usually repre
sented in the form of word embeddings and compositional functions, respectively.
 For the latter, Recurrent Neural Networks (RNNs) and Convolutional Neural Netwo
rks (CNNs) have been considered. There has not been a rigorous evaluation regard
ing the relative importance of each component to different text-representation-b
ased tasks; i.e., how important is the modeling capacity of word embeddings alon
e, relative to the added value of a compositional function? In this paper, we co
nduct an extensive comparative study between Simple Word Embeddings-based Models
 (SWEMs), with no compositional parameters, relative to employing word embedding
s within RNN/CNN-based models. Surprisingly, SWEMs exhibit comparable or even su
perior performance in the majority of cases considered. Moreover, in a new SWEM
setup, we propose to employ a max-pooling operation over the learned word-embedd
ing matrix of a given sentence. This approach is demonstrated to extract complem
entary features relative to the averaging operation standard to SWEMs, while end
owing our model with better interpretability. To further validate our observatio
ns, we examine the information utilized by different models to make predictions,
 revealing interesting properties of word embeddings.

****
SCAN: Learning Hierarchical Compositional Visual Concepts
Irina Higgins,Nicolas Sonnerat,Loic Matthey,Arka Pal,Christopher P Burgess,Matko
 Bošnjak,Murray Shanahan,Matthew Botvinick,Demis Hassabis,Alexander Lerchner
The seemingly infinite diversity of the natural world arises from a relatively s
mall set of coherent rules, such as the laws of physics or chemistry. We conject
ure that these rules give rise to regularities that can be discovered through pr
imarily unsupervised experiences and represented as abstract concepts. If such r
epresentations are compositional and hierarchical, they can be recombined into a
n exponentially large set of new concepts. This paper describes SCAN (Symbol-Con
cept Association Network), a new framework for learning such abstractions in the
 visual domain. SCAN learns concepts through fast symbol association, grounding
them in disentangled visual primitives that are discovered in an unsupervised ma
nner. Unlike state of the art multimodal generative model baselines, our approac
h requires very few pairings between symbols and images and makes no assumptions
 about the form of symbol representations. Once trained, SCAN is capable of mult
imodal bi-directional inference, generating a diverse set of image samples from
symbolic descriptions and vice versa. It also allows for traversal and manipulat
ion of the implicit hierarchy of visual concepts through symbolic instructions a
nd learnt logical recombination operations. Such manipulations enable SCAN to br
eak away from its training data distribution and imagine novel visual concepts t
hrough symbolically instructed recombination of previously learnt concepts.
****
Efficient Sparse-Winograd Convolutional Neural Networks

Xingyu Liu,Jeff Pool,Song Han,William J. Dally
Convolutional Neural Networks (CNNs) are computationally intensive, which limits their application on mobile devices. Their energy is dominated by the number of multiplies needed to perform the convolutions. Winograd’s minimal filtering algorithm (Lavin, 2015) and network pruning (Han et al., 2015) can reduce the operation count, but these two methods cannot be straightforwardly combined — applying the Winograd transform fills in the sparsity in both the weights and the activations. We propose two modifications to Winograd-based CNNs to enable these methods to exploit sparsity. First, we move the ReLU operation into the Winograd domain to increase the sparsity of the transformed activations. Second, we prune the weights in the Winograd domain to exploit static weight sparsity. For models on CIFAR-10, CIFAR-100 and ImageNet datasets, our method reduces the number of multiplications by 10.4x, 6.8x and 10.8x respectively with loss of accuracy less than 0.1%, outperforming previous baselines by 2.0x-3.0x. We also show that moving ReLU to the Winograd domain allows more aggressive pruning.
****
An inference-based policy gradient method for learning options
Matthew J. A. Smith,Herke van Hoof,Joelle Pineau
In the pursuit of increasingly intelligent learning systems, abstraction plays a vital role in enabling sophisticated decisions to be made in complex environments. The options framework provides formalism for such abstraction over sequences of decisions. However most models require that options be given a priori, presumably specified by hand, which is neither efficient, nor scalable. Indeed, it is preferable to learn options directly from interaction with the environment. Despite several efforts, this remains a difficult problem: many approaches require access to a model of the environmental dynamics, and inferred options are often not interpretable, which limits our ability to explain the system behavior for verification or debugging purposes. In this work we develop a novel policy gradient method for the automatic learning of policies with options. This algorithm uses inference methods to simultaneously improve all of the options available to an agent, and thus can be employed in an off-policy manner, without observing option labels. Experimental results show that the options learned can be interpreted. Further, we find that the method presented here is more sample efficient than existing methods, leading to faster and more stable learning of policies with options.
****
mixup: Beyond Empirical Risk Minimization
Hongyi Zhang,Moustapha Cisse,Yann N. Dauphin,David Lopez-Paz
Large deep neural networks are powerful, but exhibit undesirable behaviors such as memorization and sensitivity to adversarial examples. In this work, we propose mixup, a simple learning principle to alleviate these issues. In essence, mixup trains a neural network on convex combinations of pairs of examples and their labels. By doing so, mixup regularizes the neural network to favor simple linear behavior in-between training examples. Our experiments on the ImageNet-2012, CIFAR-10, CIFAR-100, Google commands and UCI datasets show that mixup improves the generalization of state-of-the-art neural network architectures. We also find that mixup reduces the memorization of corrupt labels, increases the robustness to adversarial examples, and stabilizes the training of generative adversarial networks.
****
Learning Deep Models: Critical Points and Local Openness
Maher Nouiehed,Meisam Razaviyayn
With the increasing interest in deeper understanding of the loss surface of many non-convex deep models, this paper presents a unifying framework to study the local/global optima equivalence of the optimization problems arising from training of such non-convex models. Using the "local openness" property of the underlying training models, we provide simple sufficient conditions under which any local optimum of the resulting optimization problem is globally optimal. We first completely characterize the local openness of matrix multiplication mapping in its range. Then we use our characterization to: 1) show that every local optimu

m of two layer linear networks is globally optimal. Unlike many existing result s in the literature, our result requires no assumption on the target data matri x Y, and input data matrix X. 2) develop almost complete characterization of the local/global optima equivalence of multi-layer linear neural networks. We provi de various counterexamples to show the necessity of each of our assumptions. 3) show global/local optima equivalence of non-linear deep models having certain py ramidal structure. Unlike some existing works, our result requires no assumption on the differentiability of the activation functions and can go beyond "full-ra nk" cases.

****

Discovering the mechanics of hidden neurons

Simon Carbonnelle,Christophe De Vleeschouwer

Neural networks trained through stochastic gradient descent (SGD) have been arou nd for more than 30 years, but they still escape our understanding. This paper t akes an experimental approach, with a divide-and-conquer strategy in mind: we st art by studying what happens in single neurons. While being the core building bl ock of deep neural networks, the way they encode information about the inputs an d how such encodings emerge is still unknown. We report experiments providing st rong evidence that hidden neurons behave like binary classifiers during training and testing. During training, analysis of the gradients reveals that a neuron s eparates two categories of inputs, which are impressively constant across traini ng. During testing, we show that the fuzzy, binary partition described above emb eds the core information used by the network for its prediction. These observati ons bring to light some of the core internal mechanics of deep neural networks, and have the potential to guide the next theoretical and practical developments.

****

Towards Neural Phrase-based Machine Translation

Po-Sen Huang,Chong Wang,Sitao Huang,Dengyong Zhou,Li Deng

In this paper, we present Neural Phrase-based Machine Translation (NPMT). Our me thod explicitly models the phrase structures in output sequences using Sleep-WAk e Networks (SWAN), a recently proposed segmentation-based sequence modeling meth od. To mitigate the monotonic alignment requirement of SWAN, we introduce a new layer to perform (soft) local reordering of input sequences. Different from exis ting neural machine translation (NMT) approaches, NPMT does not use attention-ba sed decoding mechanisms. Instead, it directly outputs phrases in a sequential o rder and can decode in linear time. Our experiments show that NPMT achieves supe rior performances on IWSLT 2014 German-English/English-German and IWSLT 2015 Eng lish-Vietnamese machine translation tasks compared with strong NMT baselines. We also observe that our method produces meaningful phrases in output languages.

****

cGANs with Projection Discriminator

Takeru Miyato,Masanori Koyama

We propose a novel, projection based way to incorporate the conditional informat ion into the discriminator of GANs that respects the role of the conditional inf ormation in the underlining probabilistic model.

This approach is in contrast with most frameworks of conditional GANs used in ap plication today, which use the conditional information by concatenating the (emb edded) conditional vector to the feature vectors.

With this modification, we were able to significantly improve the quality of the class conditional image generation on ILSVRC2012 (ImageNet) dataset from the cu rrent state-of-the-art result, and we achieved this with a single pair of a disc riminator and a generator.

We were also able to extend the application to super-resolution and succeeded in producing highly discriminative super-resolution images.

This new structure also enabled high quality category transformation based on pa rametric functional transformation of conditional batch normalization layers in the generator.

****

Multi-Task Learning for Document Ranking and Query Suggestion

Wasi Uddin Ahmad,Kai-Wei Chang,Hongning Wang
We propose a multi-task learning framework to jointly learn document ranking and query suggestion for web search. It consists of two major components, a document ranker, and a query recommender. Document ranker combines current query and session information and compares the combined representation with document representation to rank the documents. Query recommender tracks users' query reformulation sequence considering all previous in-session queries using a sequence to sequence approach. As both tasks are driven by the users' underlying search intent, we perform joint learning of these two components through session recurrence, which encodes search context and intent. Extensive comparisons against state-of-the-art document ranking and query suggestion algorithms are performed on the public AOL search log, and the promising results endorse the effectiveness of the joint learning framework.
****
Generative Adversarial Networks using Adaptive Convolution
Nhat M. Nguyen,Nilanjan Ray
Most existing GANs architectures that generate images use transposed convolution or resize-convolution as their upsampling algorithm from lower to higher resolution feature maps in the generator. We argue that this kind of fixed operation is problematic for GANs to model objects that have very different visual appearances. We propose a novel adaptive convolution method that learns the upsampling algorithm based on the local context at each location to address this problem. We modify a baseline GANs architecture by replacing normal convolutions with adaptive convolutions in the generator. Experiments on CIFAR-10 dataset show that our modified models improve the baseline model by a large margin. Furthermore, our models achieve state-of-the-art performance on CIFAR-10 and STL-10 datasets in the unsupervised setting.
****
Flipout: Efficient Pseudo-Independent Weight Perturbations on Mini-Batches
Yeming Wen,Paul Vicol,Jimmy Ba,Dustin Tran,Roger Grosse
Stochastic neural net weights are used in a variety of contexts, including regularization, Bayesian neural nets, exploration in reinforcement learning, and evolution strategies. Unfortunately, due to the large number of weights, all the examples in a mini-batch typically share the same weight perturbation, thereby limiting the variance reduction effect of large mini-batches. We introduce flipout, an efficient method for decorrelating the gradients within a mini-batch by implicitly sampling pseudo-independent weight perturbations for each example. Empirically, flipout achieves the ideal linear variance reduction for fully connected networks, convolutional networks, and RNNs. We find significant speedups in training neural networks with multiplicative Gaussian perturbations. We show that flipout is effective at regularizing LSTMs, and outperforms previous methods. Flipout also enables us to vectorize evolution strategies: in our experiments, a single GPU with flipout can handle the same throughput as at least 40 CPU cores using existing methods, equivalent to a factor-of-4 cost reduction on Amazon Web Services.
****
Improving GAN Training via Binarized Representation Entropy (BRE) Regularization
Yanshuai Cao,Gavin Weiguang Ding,Kry Yik-Chau Lui,Ruitong Huang
We propose a novel regularizer to improve the training of Generative Adversarial Networks (GANs). The motivation is that when the discriminator D spreads out its model capacity in the right way, the learning signals given to the generator G are more informative and diverse, which helps G to explore better and discover the real data manifold while avoiding large unstable jumps due to the erroneous extrapolation made by D . Our regularizer guides the rectifier discriminator D to better allocate its model capacity, by encouraging the binary activation patterns on selected internal layers of D to have a high joint entropy. Experimental results on both synthetic data and real datasets demonstrate improvements in stability and convergence speed of the GAN training, as well as higher sample quality. The approach also leads to higher classification accuracies in semi-supervised learning.

****
TD Learning with Constrained Gradients
Ishan Durugkar,Peter Stone
Temporal Difference Learning with function approximation is known to be unstable
. Previous work like \citet{sutton2009fast} and \citet{sutton2009convergent} has
 presented alternative objectives that are stable to minimize. However, in pract
ice, TD-learning with neural networks requires various tricks like using a targe
t network that updates slowly \citep{mnih2015human}. In this work we propose a c
onstraint on the TD update that minimizes change to the target values. This cons
traint can be applied to the gradients of any TD objective, and can be easily ap
plied to nonlinear function approximation. We validate this update by applying o
ur technique to deep Q-learning, and training without a target network. We also
show that adding this constraint on Baird's counterexample keeps Q-learning from
 diverging.
****
Generalizing Hamiltonian Monte Carlo with Neural Networks
Daniel Levy,Matt D. Hoffman,Jascha Sohl-Dickstein
We present a general-purpose method to train Markov chain Monte Carlo kernels, p
arameterized by deep neural networks, that converge and mix quickly to their tar
get distribution. Our method generalizes Hamiltonian Monte Carlo and is trained
to maximize expected squared jumped distance, a proxy for mixing speed. We demon
strate large empirical gains on a collection of simple but challenging distribut
ions, for instance achieving a 106x improvement in effective sample size in one
case, and mixing when standard HMC makes no measurable progress in a second. Fin
ally, we show quantitative and qualitative gains on a real-world task: latent-va
riable generative modeling. Python source code will be open-sourced with the cam
era-ready paper.
****
A Neural Method for Goal-Oriented Dialog Systems to interact with Named Entities
Janarthanan Rajendran,Jatin Ganhotra,Xiaoxiao Guo,Mo Yu,Satinder Singh
Many goal-oriented dialog tasks, especially ones in which the dialog system has
to interact with external knowledge sources such as databases, have to handle a
large number of Named Entities (NEs). There are at least two challenges in handl
ing NEs using neural methods in such settings: individual NEs may occur only rar
ely making it hard to learn good representations of them, and many of the Out Of
 Vocabulary words that occur during test time may be NEs. Thus, the need to inte
ract well with these NEs has emerged as a serious challenge to building neural m
ethods for goal-oriented dialog tasks. In this paper, we propose a new neural me
thod for this problem, and present empirical evaluations on a structured Questio
n answering task and three related goal-oriented dialog tasks that show that our
 proposed method can be effective in interacting with NEs in these settings.
****
On Optimality Conditions for Auto-Encoder Signal Recovery
Devansh Arpit,Yingbo Zhou,Hung Q. Ngo,Nils Napp,Venu Govindaraju
Auto-Encoders are unsupervised models that aim to learn patterns from observed
data by minimizing a reconstruction cost. The useful representations learned are
 often found to be sparse and distributed. On the other hand, compressed sensing
 and sparse coding assume a data generating process, where the observed data is
generated from some true latent signal source, and try to recover the correspond
ing signal from measurements. Looking at auto-encoders from this signal recovery
 perspective enables us to have a more coherent view of these techniques. In thi
s paper, in particular, we show that the true hidden representation can be appro
ximately recovered if the weight matrices are highly incoherent with unit $ \ell
^{2} $ row length and the bias vectors takes the value (approximately) equal to
the negative of the data mean. The recovery also becomes more and more accurate
as the sparsity in hidden signals increases. Additionally, we empirically also d
emonstrate that auto-encoders are capable of recovering the data generating dict
ionary when only data samples are given.
****

Tree2Tree Learning with Memory Unit
Ning Miao,Hengliang Wang,Ran Le,Chongyang Tao,Mingyue Shang,Rui Yan,Dongyan Zhao
Traditional recurrent neural network (RNN) or convolutional neural net- work (CNN) based sequence-to-sequence model can not handle tree structural data well. To alleviate this problem, in this paper, we propose a tree-to-tree model with specially designed encoder unit and decoder unit, which recursively encodes tree inputs into highly folded tree embeddings and decodes the embeddings into tree outputs. Our model could represent the complex information of a tree while also restore a tree from embeddings.
We evaluate our model in random tree recovery task and neural machine translation task. Experiments show that our model outperforms the baseline model.
****
Convergence rate of sign stochastic gradient descent for non-convex functions
Jeremy Bernstein,Kamyar Azizzadenesheli,Yu-Xiang Wang,Anima Anandkumar
The sign stochastic gradient descent method (signSGD) utilizes only the sign of the stochastic gradient in its updates. Since signSGD carries out one-bit quantization of the gradients, it is extremely practical for distributed optimization where gradients need to be aggregated from different processors. For the first time, we establish convergence rates for signSGD on general non-convex functions under transparent conditions. We show that the rate of signSGD to reach first-order critical points matches that of SGD in terms of number of stochastic gradient calls, up to roughly a linear factor in the dimension. We carry out simple experiments to explore the behaviour of sign gradient descent (without the stochasticity) close to saddle points and show that it often helps completely avoid them without using either stochasticity or curvature information.
****
Improving the Universality and Learnability of Neural Programmer-Interpreters with Combinator Abstraction
Da Xiao,Jo-Yu Liao,Xingyuan Yuan
To overcome the limitations of Neural Programmer-Interpreters (NPI) in its universality and learnability, we propose the incorporation of combinator abstraction into neural programing and a new NPI architecture to support this abstraction, which we call Combinatory Neural Programmer-Interpreter (CNPI). Combinator abstraction dramatically reduces the number and complexity of programs that need to be interpreted by the core controller of CNPI, while still allowing the CNPI to represent and interpret arbitrary complex programs by the collaboration of the core with the other components. We propose a small set of four combinators to capture the most pervasive programming patterns. Due to the finiteness and simplicity of this combinator set and the offloading of some burden of interpretation from the core, we are able construct a CNPI that is universal with respect to the set of all combinatorizable programs, which is adequate for solving most algorithmic tasks. Moreover, besides supervised training on execution traces, CNPI can be trained by policy gradient reinforcement learning with appropriately designed curricula.
****
Minimax Curriculum Learning: Machine Teaching with Desirable Difficulties and Scheduled Diversity
Tianyi Zhou,Jeff Bilmes
We introduce and study minimax curriculum learning (MCL), a new method for adaptively selecting a sequence of training subsets for a succession of stages in machine learning. The subsets are encouraged to be small and diverse early on, and then larger, harder, and allowably more homogeneous in later stages. At each stage, model weights and training sets are chosen by solving a joint continuous-discrete minimax optimization, whose objective is composed of a continuous loss (reflecting training set hardness) and a discrete submodular promoter of diversity for the chosen subset. MCL repeatedly solves a sequence of such optimizations with a schedule of increasing training set size and decreasing pressure on diversity encouragement. We reduce MCL to the minimization of a surrogate function handled by submodular maximization and continuous gradient methods. We show that MCL achieves better performance and, with a clustering trick, uses fewer labeled sa

mples for both shallow and deep models while achieving the same performance. Our method involves repeatedly solving constrained submodular maximization of an only slowly varying function on the same ground set. Therefore, we develop a heuristic method that utilizes the previous submodular maximization solution as a warm start for the current submodular maximization process to reduce computation while still yielding a guarantee.
****

Deep Lipschitz networks and Dudley GANs

Ehsan Abbasnejad,Javen Shi,Anton van den Hengel

Generative adversarial networks (GANs) have enjoyed great success, however often suffer instability during training which motivates many attempts to resolve this issue. Theoretical explanation for the cause of instability is provided in Wasserstein GAN (WGAN), and wasserstein distance is proposed to stablize the training. Though WGAN is indeed more stable than previous GANs, it takes much more iterations and time to train. This is because the ways to ensure Lipschitz condition in WGAN (such as weight-clipping) significantly limit the capacity of the network. In this paper, we argue that it is beneficial to ensure Lipschitz condition as well as maintain sufficient capacity and expressiveness of the network. To facilitate this, we develop both theoretical and practical building blocks, using which one can construct different neural networks using a large range of metrics, as well as ensure Lipschitz condition and sufficient capacity of the networks. Using the proposed building blocks, and a special choice of a metric called Dudley metric, we propose Dudley GAN that outperforms the state of the arts in both convergence and sample quality. We discover a natural link between Dudley GAN (and its extension) and empirical risk minimization, which gives rise to generalization analysis.
****

Tensor Contraction & Regression Networks

Jean Kossaifi,Zack Chase Lipton,Aran Khanna,Tommaso Furlanello,Anima Anandkumar

Convolution neural networks typically consist of many convolutional layers followed by several fully-connected layers. While convolutional layers map between high-order activation tensors, the fully-connected layers operate on flattened activation vectors. Despite its success, this approach has notable drawbacks. Flattening discards the multi-dimensional structure of the activations, and the fully-connected layers require a large number of parameters.

We present two new techniques to address these problems. First, we introduce tensor contraction layers which can replace the ordinary fully-connected layers in a neural network. Second, we introduce tensor regression layers, which express the output of a neural network as a low-rank multi-linear mapping from a high-order activation tensor to the softmax layer. Both the contraction and regression weights are learned end-to-end by backpropagation. By imposing low rank on both, we use significantly fewer parameters. Experiments on the ImageNet dataset show that applied to the popular VGG and ResNet architectures, our methods significantly reduce the number of parameters in the fully connected layers (about 65% space savings) while negligibly impacting accuracy.
****

Continuous Adaptation via Meta-Learning in Nonstationary and Competitive Environments

Maruan Al-Shedivat,Trapit Bansal,Yura Burda,Ilya Sutskever,Igor Mordatch,Pieter Abbeel

Ability to continuously learn and adapt from limited experience in nonstationary environments is an important milestone on the path towards general intelligence. In this paper, we cast the problem of continuous adaptation into the learning-to-learn framework. We develop a simple gradient-based meta-learning algorithm suitable for adaptation in dynamically changing and adversarial scenarios. Additionally, we design a new multi-agent competitive environment, RoboSumo, and define iterated adaptation games for testing various aspects of continuous adaptation. We demonstrate that meta-learning enables significantly more efficient adaptation than reactive baselines in the few-shot regime. Our experiments with a population of agents that learn and compete suggest that meta-learners are the fittes

t.
****
Value Propagation Networks
Nantas Nardelli,Gabriel Synnaeve,Zeming Lin,Pushmeet Kohli,Nicolas Usunier
We present Value Propagation (VProp), a parameter-efficient differentiable planning module built on Value Iteration which can successfully be trained in a reinforcement learning fashion to solve unseen tasks, has the capability to generalize to larger map sizes, and can learn to navigate in dynamic environments. We evaluate on configurations of MazeBase grid-worlds, with randomly generated environments of several different sizes. Furthermore, we show that the module enables to learn to plan when the environment also includes stochastic elements, providing a cost-efficient learning system to build low-level size-invariant planners for a variety of interactive navigation problems.
****
Reward Estimation via State Prediction
Daiki Kimura,Subhajit Chaudhury,Ryuki Tachibana,Sakyasingha Dasgupta
Reinforcement learning typically requires carefully designed reward functions in order to learn the desired behavior. We present a novel reward estimation method that is based on a finite sample of optimal state trajectories from expert demon- strations and can be used for guiding an agent to mimic the expert behavior. The optimal state trajectories are used to learn a generative or predictive model of the “good” states distribution. The reward signal is computed by a function of the difference between the actual next state acquired by the agent and the predicted next state given by the learned generative or predictive model. With this inferred reward function, we perform standard reinforcement learning in the inner loop to guide the agent to learn the given task. Experimental evaluations across a range of tasks demonstrate that the proposed method produces superior performance compared to standard reinforcement learning with both complete or sparse hand engineered rewards. Furthermore, we show that our method successfully enables an agent to learn good actions directly from expert player video of games such as the Super Mario Bros and Flappy Bird.
****
Distributed non-parametric deep and wide networks
Biswa Sengupta,Yu Qian
In recent work, it was shown that combining multi-kernel based support vector machines (SVMs) can lead to near state-of-the-art performance on an action recognition dataset (HMDB-51 dataset). In the present work, we show that combining distributed Gaussian Processes with multi-stream deep convolutional neural networks (CNN) alleviate the need to augment a neural network with hand-crafted features. In contrast to prior work, we treat each deep neural convolutional network as an expert wherein the individual predictions (and their respective uncertainties) are combined into a Product of Experts (PoE) framework.
****
An Out-of-the-box Full-network Embedding for Convolutional Neural Networks
Dario Garcia-Gasulla,Armand Vilalta,Ferran Parés,Jonatan Moreno,Eduard Ayguadé,Jesús Labarta,Ulises Cortés,Toyotaro Suzumura
Transfer learning for feature extraction can be used to exploit deep representations in contexts where there is very few training data, where there are limited computational resources, or when tuning the hyper-parameters needed for training is not an option. While previous contributions to feature extraction propose embeddings based on a single layer of the network, in this paper we propose a full-network embedding which successfully integrates convolutional and fully connected features, coming from all layers of a deep convolutional neural network. To do so, the embedding normalizes features in the context of the problem, and discretizes their values to reduce noise and regularize the embedding space. Significantly, this also reduces the computational cost of processing the resultant representations. The proposed method is shown to outperform single layer embeddings on several image classification tasks, while also being more robust to the choice of the pre-trained model used for obtaining the initial features. The performance gap in classification accuracy between thoroughly tuned solutions and the fu

ll-network embedding is also reduced, which makes of the proposed approach a competitive solution for a large set of applications.
****

Recasting Gradient-Based Meta-Learning as Hierarchical Bayes
Erin Grant,Chelsea Finn,Sergey Levine,Trevor Darrell,Thomas Griffiths
Meta-learning allows an intelligent agent to leverage prior learning episodes as a basis for quickly improving performance on a novel task. Bayesian hierarchical modeling provides a theoretical framework for formalizing meta-learning as inference for a set of parameters that are shared across tasks. Here, we reformulate the model-agnostic meta-learning algorithm (MAML) of Finn et al. (2017) as a method for probabilistic inference in a hierarchical Bayesian model. In contrast to prior methods for meta-learning via hierarchical Bayes, MAML is naturally applicable to complex function approximators through its use of a scalable gradient descent procedure for posterior inference. Furthermore, the identification of MAML as hierarchical Bayes provides a way to understand the algorithm’s operation as a meta-learning procedure, as well as an opportunity to make use of computational strategies for efficient inference. We use this opportunity to propose an improvement to the MAML algorithm that makes use of techniques from approximate inference and curvature estimation.
****

Decoupling the Layers in Residual Networks
Ricky Fok,Aijun An,Zana Rashidi,Xiaogang Wang
We propose a Warped Residual Network (WarpNet) using a parallelizable warp operator for forward and backward propagation to distant layers that trains faster than the original residual neural network. We apply a perturbation theory on residual networks and decouple the interactions between residual units. The resulting warp operator is a first order approximation of the output over multiple layers. The first order perturbation theory exhibits properties such as binomial path lengths and exponential gradient scaling found experimentally by Veit et al (2016).
We demonstrate through an extensive performance study that the proposed network achieves comparable predictive performance to the original residual network with the same number of parameters, while achieving a significant speed-up on the total training time. As WarpNet performs model parallelism in residual network training in which weights are distributed over different GPUs, it offers speed-up and capability to train larger networks compared to original residual networks.
****

Proximal Backpropagation
Thomas Frerix,Thomas Möllenhoff,Michael Moeller,Daniel Cremers
We propose proximal backpropagation (ProxProp) as a novel algorithm that takes implicit instead of explicit gradient steps to update the network parameters during neural network training. Our algorithm is motivated by the step size limitation of explicit gradient descent, which poses an impediment for optimization. ProxProp is developed from a general point of view on the backpropagation algorithm, currently the most common technique to train neural networks via stochastic gradient descent and variants thereof. Specifically, we show that backpropagation of a prediction error is equivalent to sequential gradient descent steps on a quadratic penalty energy, which comprises the network activations as variables of the optimization. We further analyze theoretical properties of ProxProp and in particular prove that the algorithm yields a descent direction in parameter space and can therefore be combined with a wide variety of convergent algorithms. Finally, we devise an efficient numerical implementation that integrates well with popular deep learning frameworks. We conclude by demonstrating promising numerical results and show that ProxProp can be effectively combined with common first order optimizers such as Adam.
****

Neuron as an Agent
Shohei Ohsawa,Kei Akuzawa,Tatsuya Matsushima,Gustavo Bezerra,Yusuke Iwasawa,Hiroshi Kajino,Seiya Takenaka,Yutaka Matsuo
Existing multi-agent reinforcement learning (MARL) communication methods have re

lied on a trusted third party (TTP) to distribute reward to agents, leaving them inapplicable in peer-to-peer environments. This paper proposes reward distribution using {\em Neuron as an Agent} (NaaA) in MARL without a TTP with two key ideas: (i) inter-agent reward distribution and (ii) auction theory. Auction theory is introduced because inter-agent reward distribution is insufficient for optimization. Agents in NaaA maximize their profits (the difference between reward and cost) and, as a theoretical result, the auction mechanism is shown to have agents autonomously evaluate counterfactual returns as the values of other agents. NaaA enables representation trades in peer-to-peer environments, ultimately regarding unit in neural networks as agents. Finally, numerical experiments (a single-agent environment from OpenAI Gym and a multi-agent environment from ViZDoom) confirm that NaaA framework optimization leads to better performance in reinforcement learning.
****
Bounding and Counting Linear Regions of Deep Neural Networks
Thiago Serra,Christian Tjandraatmadja,Srikumar Ramalingam
In this paper, we study the representational power of deep neural networks (DNN) that belong to the family of piecewise-linear (PWL) functions, based on PWL activation units such as rectifier or maxout. We investigate the complexity of such networks by studying the number of linear regions of the PWL function. Typically, a PWL function from a DNN can be seen as a large family of linear functions acting on millions of such regions. We directly build upon the work of Mont´ufar et al. (2014), Mont´ufar (2017), and Raghu et al. (2017) by refining the upper and lower bounds on the number of linear regions for rectified and maxout networks. In addition to achieving tighter bounds, we also develop a novel method to perform exact numeration or counting of the number of linear regions with a mixed-integer linear formulation that maps the input space to output. We use this new capability to visualize how the number of linear regions change while training DNNs.
****
ElimiNet: A Model for Eliminating Options for Reading Comprehension with Multiple Choice Questions
Soham Parikh,Ananya Sai,Preksha Nema,Mitesh M Khapra
The task of Reading Comprehension with Multiple Choice Questions, requires a human (or machine) to read a given \{\textit{passage, question}\} pair and select one of the n given options. The current state of the art model for this task first computes a query-aware representation for the passage and then \textit{selects} the option which has the maximum similarity with this representation. However, when humans perform this task they do not just focus on option selection but use a combination of \textit{elimination} and \textit{selection}. Specifically, a human would first try to eliminate the most irrelevant option and then read the document again in the light of this new information (and perhaps ignore portions corresponding to the eliminated option). This process could be repeated multiple times till the reader is finally ready to select the correct option. We propose \textit{ElimiNet}, a neural network based model which tries to mimic this process. Specifically, it has gates which decide whether an option can be eliminated given the \{\textit{document, question}\} pair and if so it tries to make the document representation orthogonal to this eliminatedd option (akin to ignoring portions of the document corresponding to the eliminated option). The model makes multiple rounds of partial elimination to refine the document representation and finally uses a selection module to pick the best option. We evaluate our model on the recently released large scale RACE dataset and show that it outperforms the current state of the art model on 7 out of the 13 question types in this dataset. Further we show that taking an ensemble of our \textit{elimination-selection} based method with a \textit{selection} based method gives us an improvement of 7\% (relative) over the best reported performance on this dataset.

****
Learning to search with MCTSnets
Arthur Guez,Theophane Weber,Ioannis Antonoglou,Karen Simonyan,Oriol Vinyals,Daan

Wierstra,Remi Munos,David Silver
Planning problems are among the most important and well-studied problems in arti ficial intelligence. They are most typically solved by tree search algorithms th at simulate ahead into the future, evaluate future states, and back-up those eva luations to the root of a search tree. Among these algorithms, Monte-Carlo tree search (MCTS) is one of the most general, powerful and widely used. A typical im plementation of MCTS uses cleverly designed rules, optimised to the particular c haracteristics of the domain. These rules control where the simulation traverses , what to evaluate in the states that are reached, and how to back-up those eval uations. In this paper we instead learn where, what and how to search. Our archi tecture, which we call an MCTSnet, incorporates simulation-based search inside a neural network, by expanding, evaluating and backing-up a vector embedding. The parameters of the network are trained end-to-end using gradient-based optimisat ion. When applied to small searches in the well-known planning problem Sokoban, the learned search algorithm significantly outperformed MCTS baselines.
****
An Online Learning Approach to Generative Adversarial Networks
Paulina Grnarova,Kfir Y Levy,Aurelien Lucchi,Thomas Hofmann,Andreas Krause
We consider the problem of training generative models with a Generative Adversar ial Network (GAN). Although GANs can accurately model complex distributions, the y are known to be difficult to train due to instabilities caused by a difficult minimax optimization problem. In this paper, we view the problem of training GAN s as finding a mixed strategy in a zero-sum game. Building on ideas from online learning we propose a novel training method named Chekhov GAN. On the theory sid e, we show that our method provably converges to an equilibrium for semi-shallow GAN architectures, i.e. architectures where the discriminator is a one-layer ne twork and the generator is arbitrary. On the practical side, we develop an effic ient heuristic guided by our theoretical results, which we apply to commonly use d deep GAN architectures.
On several real-world tasks our approach exhibits improved stability and perform ance compared to standard GAN training.
****
No Spurious Local Minima in a Two Hidden Unit ReLU Network
Chenwei Wu,Jiajun Luo,Jason D. Lee
Deep learning models can be efficiently optimized via stochastic gradient descen t, but there is little theoretical evidence to support this. A key question in o ptimization is to understand when the optimization landscape of a neural network is amenable to gradient-based optimization. We focus on a simple neural network two-layer ReLU network with two hidden units, and show that all local minimizer s are global. This combined with recent work of Lee et al. (2017); Lee et al. (2 016) show that gradient descent converges to the global minimizer.
****
Unsupervised Adversarial Anomaly Detection using One-Class Support Vector Machi nes
Prameesha Sandamal Weerasinghe,Tansu Alpcan,Sarah Monazam Erfani,Christopher Lec kie
Anomaly detection discovers regular patterns in unlabeled data and identifies th e non-conforming data points, which in some cases are the result of malicious at tacks by adversaries. Learners such as One-Class Support Vector Machines (OCSVMs) have been successfully in anomaly detection, yet their performance may degrade significantly in the presence of sophisticated adversaries, who target the algo rithm itself by compromising the integrity of the training data. With the rise i n the use of machine learning in mission critical day-to-day activities where er rors may have significant consequences, it is imperative that machine learning s ystems are made secure. To address this, we propose a defense mechanism that is based on a contraction of the data, and we test its effectiveness using OCSVMs. The proposed approach introduces a layer of uncertainty on top of the OCSVM lear ner, making it infeasible for the adversary to guess the specific configuration of the learner. We theoretically analyze the effects of adversarial perturbation s on the separating margin of OCSVMs and provide empirical evidence on several b

enchmark datasets, which show that by carefully contracting the data in low dimensional spaces, we can successfully identify adversarial samples that would not have been identifiable in the original dimensional space. The numerical results show that the proposed method improves OCSVMs performance significantly (2-7%)
****

Deep Gradient Compression: Reducing the Communication Bandwidth for Distributed Training

Yujun Lin,Song Han,Huizi Mao,Yu Wang,Bill Dally

Large-scale distributed training requires significant communication bandwidth for gradient exchange that limits the scalability of multi-node training, and requires expensive high-bandwidth network infrastructure. The situation gets even worse with distributed training on mobile devices (federated learning), which suffers from higher latency, lower throughput, and intermittent poor connections. In this paper, we find 99.9% of the gradient exchange in distributed SGD is redundant, and propose Deep Gradient Compression (DGC) to greatly reduce the communication bandwidth. To preserve accuracy during compression, DGC employs four methods: momentum correction, local gradient clipping, momentum factor masking, and warm-up training. We have applied Deep Gradient Compression to image classification, speech recognition, and language modeling with multiple datasets including Cifar10, ImageNet, Penn Treebank, and Librispeech Corpus. On these scenarios, Deep Gradient Compression achieves a gradient compression ratio from 270x to 600x without losing accuracy, cutting the gradient size of ResNet-50 from 97MB to 0.35MB, and for DeepSpeech from 488MB to 0.74MB. Deep gradient compression enables large-scale distributed training on inexpensive commodity 1Gbps Ethernet and facilitates distributed training on mobile.
****

Mixed Precision Training of Convolutional Neural Networks using Integer Operations

Dipankar Das,Naveen Mellempudi,Dheevatsa Mudigere,Dhiraj Kalamkar,Sasikanth Avancha,Kunal Banerjee,Srinivas Sridharan,Karthik Vaidyanathan,Bharat Kaul,Evangelos Georganas,Alexander Heinecke,Pradeep Dubey,Jesus Corbal,Nikita Shustrov,Roma Dubtsov,Evarist Fomenko,Vadim Pirogov

The state-of-the-art (SOTA) for mixed precision training is dominated by variants of low precision floating point operations, and in particular, FP16 accumulating into FP32 Micikevicius et al. (2017). On the other hand, while a lot of research has also happened in the domain of low and mixed-precision Integer training, these works either present results for non-SOTA networks (for instance only AlexNet for ImageNet-1K), or relatively small datasets (like CIFAR-10). In this work, we train state-of-the-art visual understanding neural networks on the ImageNet-1K dataset, with Integer operations on General Purpose (GP) hardware. In particular, we focus on Integer Fused-Multiply-and-Accumulate (FMA) operations which take two pairs of INT16 operands and accumulate results into an INT32 output.We propose a shared exponent representation of tensors and develop a Dynamic Fixed Point (DFP) scheme suitable for common neural network operations. The nuances of developing an efficient integer convolution kernel is examined, including methods to handle overflow of the INT32 accumulator. We implement CNN training for ResNet-50, GoogLeNet-v1, VGG-16 and AlexNet; and these networks achieve or exceed SOTA accuracy within the same number of iterations as their FP32 counterparts without any change in hyper-parameters and with a 1.8X improvement in end-to-end training throughput. To the best of our knowledge these results represent the first INT16 training results on GP hardware for ImageNet-1K dataset using SOTA CNNs and achieve highest reported accuracy using half precision
****

Gradient Estimators for Implicit Models

Yingzhen Li,Richard E. Turner

Implicit models, which allow for the generation of samples but not for point-wise evaluation of probabilities, are omnipresent in real-world problems tackled by machine learning and a hot topic of current research. Some examples include data simulators that are widely used in engineering and scientific research, generative adversarial networks (GANs) for image synthesis, and hot-off-the-press appr

oximate inference techniques relying on implicit distributions. The majority of existing approaches to learning implicit models rely on approximating the intractable distribution or optimisation objective for gradient-based optimisation, which is liable to produce inaccurate updates and thus poor models. This paper alleviates the need for such approximations by proposing the \emph{Stein gradient estimator}, which directly estimates the score function of the implicitly defined distribution. The efficacy of the proposed estimator is empirically demonstrated by examples that include meta-learning for approximate inference and entropy regularised GANs that provide improved sample diversity.
****

Sensor Transformation Attention Networks
Stefan Braun,Daniel Neil,Enea Ceolini,Jithendar Anumula,Shih-Chii Liu
Recent work on encoder-decoder models for sequence-to-sequence mapping has shown that integrating both temporal and spatial attentional mechanisms into neural networks increases the performance of the system substantially. We report on a new modular network architecture that applies an attentional mechanism not on temporal and spatial regions of the input, but on sensor selection for multi-sensor setups. This network called the sensor transformation attention network (STAN) is evaluated in scenarios which include the presence of natural noise or synthetic dynamic noise. We demonstrate how the attentional signal responds dynamically to changing noise levels and sensor-specific noise, leading to reduced word error rates (WERs) on both audio and visual tasks using TIDIGITS and GRID; and also on CHiME-3, a multi-microphone real-world noisy dataset. The improvement grows as more channels are corrupted as demonstrated on the CHiME-3 dataset. Moreover, the proposed STAN architecture naturally introduces a number of advantages including ease of removing sensors from existing architectures, attentional interpretability, and increased robustness to a variety of noise environments.
****

Learning to navigate by distilling visual information and natural language instructions
Abhishek Sinha,Akilesh B,Mausoom Sarkar,Balaji Krishnamurthy
In this work, we focus on the problem of grounding language by training an agent to follow a set of natural language instructions and navigate to a target object in a 2D grid environment. The agent receives visual information through raw pixels and a natural language instruction telling what task needs to be achieved.
Other than these two sources of information, our model does not have any prior information of both the visual and textual modalities and is end-to-end trainable.
We develop an attention mechanism for multi-modal fusion of visual and textual modalities that allows the agent to learn to complete the navigation tasks and also
achieve language grounding. Our experimental results show that our attention mechanism outperforms the existing multi-modal fusion mechanisms proposed in order to solve the above mentioned navigation task. We demonstrate through the visualization of attention weights that our model learns to correlate attributes of
the object referred in the instruction with visual representations and also show that the learnt textual representations are semantically meaningful as they follow
vector arithmetic and are also consistent enough to induce translation between instructions
in different natural languages. We also show that our model generalizes effectively to unseen scenarios and exhibit zero-shot generalization capabilities.
In order to simulate the above described challenges, we introduce a new 2D environment
for an agent to jointly learn visual and textual modalities
****

Deep Autoencoding Gaussian Mixture Model for Unsupervised Anomaly Detection

Bo Zong,Qi Song,Martin Renqiang Min,Wei Cheng,Cristian Lumezanu,Daeki Cho,Haifen
g Chen
Unsupervised anomaly detection on multi- or high-dimensional data is of great im
portance in both fundamental machine learning research and industrial applicatio
ns, for which density estimation lies at the core. Although previous approaches
based on dimensionality reduction followed by density estimation have made fruit
ful progress, they mainly suffer from decoupled model learning with inconsistent
 optimization goals and incapability of preserving essential information in the
low-dimensional space. In this paper, we present a Deep Autoencoding Gaussian Mi
xture Model (DAGMM) for unsupervised anomaly detection. Our model utilizes a dee
p autoencoder to generate a low-dimensional representation and reconstruction er
ror for each input data point, which is further fed into a Gaussian Mixture Mode
l (GMM). Instead of using decoupled two-stage training and the standard Expectat
ion-Maximization (EM) algorithm, DAGMM jointly optimizes the parameters of the d
eep autoencoder and the mixture model simultaneously in an end-to-end fashion, l
everaging a separate estimation network to facilitate the parameter learning of
the mixture model. The joint optimization, which well balances autoencoding reco
nstruction, density estimation of latent representation, and regularization, hel
ps the autoencoder escape from less attractive local optima and further reduce r
econstruction errors, avoiding the need of pre-training. Experimental results on
 several public benchmark datasets show that, DAGMM significantly outperforms st
ate-of-the-art anomaly detection techniques, and achieves up to 14% improvement
based on the standard F1 score.
****
Bayesian Time Series Forecasting with Change Point and Anomaly Detection
Anderson Y. Zhang,Miao Lu,Deguang Kong,Jimmy Yang
Time series forecasting plays a crucial role in marketing, finance and many othe
r quantitative fields. A large amount of methodologies has been developed on thi
s topic, including ARIMA, Holt–Winters, etc. However, their performance is easil
y undermined by the existence of change points and anomaly points, two structure
s commonly observed in real data, but rarely considered in the aforementioned me
thods. In this paper, we propose a novel state space time series model, with the
 capability to capture the structure of change points and anomaly points, as wel
l as trend and seasonality. To infer all the hidden variables, we develop a Baye
sian framework, which is able to obtain distributions and forecasting intervals
for time series forecasting, with provable theoretical properties. For implement
ation, an iterative algorithm with Markov chain Monte Carlo (MCMC), Kalman filte
r and Kalman smoothing is proposed. In both synthetic data and real data applica
tions, our methodology yields a better performance in time series forecasting co
mpared with existing methods, along with more accurate change point detection an
d anomaly detection.
****
Robustness of Classifiers to Universal Perturbations: A Geometric Perspective
Seyed-Mohsen Moosavi-Dezfooli,Alhussein Fawzi,Omar Fawzi,Pascal Frossard,Stefano
 Soatto
Deep networks have recently been shown to be vulnerable to universal perturbatio
ns: there exist very small image-agnostic perturbations that cause most natural
images to be misclassified by such classifiers. In this paper, we provide a quan
titative analysis of the robustness of classifiers to universal perturbations, a
nd draw a formal link between the robustness to universal perturbations, and the
 geometry of the decision boundary. Specifically, we establish theoretical bound
s on the robustness of classifiers under two decision boundary models (flat and
curved models). We show in particular that the robustness of deep networks to un
iversal perturbations is driven by a key property of their curvature: there exis
t shared directions along which the decision boundary of deep networks is system
atically positively curved. Under such conditions, we prove the existence of sma
ll universal perturbations. Our analysis further provides a novel geometric meth
od for computing universal perturbations, in addition to explaining their proper
ties.
****

Training with Growing Sets: A Simple Alternative to Curriculum Learning and Self Paced Learning
Melike Nur Mermer,Mehmet Fatih Amasyali
Curriculum learning and Self paced learning are popular topics in the machine learning that suggest to put the training samples in order by considering their difficulty levels. Studies in these topics show that starting with a small training set and adding new samples according to difficulty levels improves the learning performance. In this paper we experimented that we can also obtain good results by adding the samples randomly without a meaningful order. We compared our method with classical training, Curriculum learning, Self paced learning and their reverse ordered versions. Results of the statistical tests show that the proposed method is better than classical method and similar with the others. These results point a new training regime that removes the process of difficulty level determination in Curriculum and Self paced learning and as successful as these methods.
****
Autoregressive Convolutional Neural Networks for Asynchronous Time Series
Mikolaj Binkowski,Gautier Marti,Philippe Donnat
We propose Significance-Offset Convolutional Neural Network, a deep convolutional network architecture for regression of multivariate asynchronous time series. The model is inspired by standard autoregressive (AR) models and gating mechanisms used in recurrent neural networks. It involves an AR-like weighting system, where the final predictor is obtained as a weighted sum of adjusted regressors, while the weights are data-dependent functions learnt through a convolutional network. The architecture was designed for applications on asynchronous time series and is evaluated on such datasets: a hedge fund proprietary dataset of over 2 million quotes for a credit derivative index, an artificially generated noisy autoregressive series and household electricity consumption dataset. The pro-posed architecture achieves promising results as compared to convolutional and recurrent neural networks. The code for the numerical experiments and the architecture implementation will be shared online to make the research reproducible.
****
CrescendoNet: A Simple Deep Convolutional Neural Network with Ensemble Behavior
Xiang Zhang,Nishant Vishwamitra,Hongxin Hu,Feng Luo
We introduce a new deep convolutional neural network, CrescendoNet, by stacking simple building blocks without residual connections. Each Crescendo block contains independent convolution paths with increased depths. The numbers of convolution layers and parameters are only increased linearly in Crescendo blocks. In experiments, CrescendoNet with only 15 layers outperforms almost all networks without residual connections on benchmark datasets, CIFAR10, CIFAR100, and SVHN. Given sufficient amount of data as in SVHN dataset, CrescendoNet with 15 layers and 4.1M parameters can match the performance of DenseNet-BC with 250 layers and 15.3M parameters. CrescendoNet provides a new way to construct high performance deep convolutional neural networks without residual connections. Moreover, through investigating the behavior and performance of subnetworks in CrescendoNet, we note that the high performance of CrescendoNet may come from its implicit ensemble behavior, which differs from the FractalNet that is also a deep convolutional neural network without residual connections. Furthermore, the independence between paths in CrescendoNet allows us to introduce a new path-wise training procedure, which can reduce the memory needed for training.
****
Sparse Persistent RNNs: Squeezing Large Recurrent Networks On-Chip
Feiwen Zhu,Jeff Pool,Michael Andersch,Jeremy Appleyard,Fung Xie
Recurrent Neural Networks (RNNs) are powerful tools for solving sequence-based problems, but their efficacy and execution time are dependent on the size of the network. Following recent work in simplifying these networks with model pruning and a novel mapping of work onto GPUs, we design an efficient implementation for sparse RNNs. We investigate several optimizations and tradeoffs: Lamport timestamps, wide memory loads, and a bank-aware weight layout. With these optimizat

ions, we achieve speedups of over 6x over the next best algorithm for a hidden layer of size 2304, batch size of 4, and a density of 30%. Further, our technique allows for models of over 5x the size to fit on a GPU for a speedup of 2x, enabling larger networks to help advance the state-of-the-art. We perform case studies on NMT and speech recognition tasks in the appendix, accelerating their recurrent layers by up to 3x.
****

Learning to Mix n-Step Returns: Generalizing Lambda-Returns for Deep Reinforcement Learning
Sahil Sharma,Girish Raguvir J *,Srivatsan Ramesh *,Balaraman Ravindran
Reinforcement Learning (RL) can model complex behavior policies for goal-directed sequential decision making tasks. A hallmark of RL algorithms is Temporal Difference (TD) learning: value function for the current state is moved towards a bootstrapped target that is estimated using the next state's value function. lambda-returns define the target of the RL agent as a weighted combination of rewards estimated by using multiple many-step look-aheads. Although mathematically tractable, the use of exponentially decaying weighting of n-step returns based targets in lambda-returns is a rather ad-hoc design choice. Our major contribution is that we propose a generalization of lambda-returns called Confidence-based Autodidactic Returns (CAR), wherein the RL agent learns the weighting of the n-step returns in an end-to-end manner. In contrast to lambda-returns wherein the RL agent is restricted to use an exponentially decaying weighting scheme, CAR allows the agent to learn to decide how much it wants to weigh the n-step returns based targets. Our experiments, in addition to showing the efficacy of CAR, also empirically demonstrate that using sophisticated weighted mixtures of multi-step returns (like CAR and lambda-returns) considerably outperforms the use of n-step returns. We perform our experiments on the Asynchronous Advantage Actor Critic (A3C) algorithm in the Atari 2600 domain.
****

Not-So-Random Features
Brian Bullins,Cyril Zhang,Yi Zhang
We propose a principled method for kernel learning, which relies on a Fourier-analytic characterization of translation-invariant or rotation-invariant kernels. Our method produces a sequence of feature maps, iteratively refining the SVM margin. We provide rigorous guarantees for optimality and generalization, interpreting our algorithm as online equilibrium-finding dynamics in a certain two-player min-max game. Evaluations on synthetic and real-world datasets demonstrate scalability and consistent improvements over related random features-based methods.
****

Dense Recurrent Neural Network with Attention Gate
Yong-Ho Yoo,Kook Han,Sanghyun Cho,Kyoung-Chul Koh,Jong-Hwan Kim
We propose the dense RNN, which has the fully connections from each hidden state to multiple preceding hidden states of all layers directly. As the density of the connection increases, the number of paths through which the gradient flows can be increased. It increases the magnitude of gradients, which help to prevent the vanishing gradient problem in time. Larger gradients, however, can also cause exploding gradient problem. To complement the trade-off between two problems, we propose an attention gate, which controls the amounts of gradient flows. We describe the relation between the attention gate and the gradient flows by approximation. The experiment on the language modeling using Penn Treebank corpus shows dense connections with the attention gate improve the model’s performance.
****

DOUBLY STOCHASTIC ADVERSARIAL AUTOENCODER
Mahdi Azarafrooz
Any autoencoder network can be turned into a generative model by imposing an arbitrary prior distribution on its hidden code vector. Variational Autoencoder uses a KL divergence penalty to impose the prior, whereas Adversarial Autoencoder uses generative adversarial networks. A straightforward modification of Adversarial Autoencoder can be achieved by replacing the adversarial network with maximum mean discrepancy (MMD) network. This replacement leads to a new set of probabi

listic autoencoder which is also discussed in our paper.

However, an essential challenge remains in both of these probabilistic autoencoders, namely that the only source of randomness at the output of encoder, is the training data itself. Lack of enough stochasticity can make the optimization problem non-trivial. As a result, they can lead to degenerate solutions where the generator collapses into sampling only a few modes.

Our proposal is to replace the adversary of the adversarial autoencoder by a space of {\it stochastic} functions. This replacement introduces a a new source of randomness which can be considered as a continuous control for encouraging {\it explorations}. This prevents the adversary from fitting too closely to the generator and therefore leads to more diverse set of generated samples. Consequently, the decoder serves as a better generative network which unlike MMD nets scales linearly with the amount of data. We provide mathematical and empirical evidence on how this replacement outperforms the pre-existing architectures.

****

An Ensemble of Retrieval-Based and Generation-Based Human-Computer Conversation Systems.

Yiping Song,Rui Yan,Cheng-Te Li,Jian-Yun Nie,Ming Zhang,Dongyan Zhao

Human-computer conversation systems have attracted much attention in Natural Language Processing. Conversation systems can be roughly divided into two categories: retrieval-based and generation-based systems. Retrieval systems search a user-issued utterance (namely a query) in a large conversational repository and return a reply that best matches the query. Generative approaches synthesize new replies. Both ways have certain advantages but suffer from their own disadvantages. We propose a novel ensemble of retrieval-based and generation-based conversation system. The retrieved candidates, in addition to the original query, are fed to a reply generator via a neural network, so that the model is aware of more information. The generated reply together with the retrieved ones then participates in a re-ranking process to find the final reply to output. Experimental results show that such an ensemble system outperforms each single module by a large margin.

****

GraphGAN: Generating Graphs via Random Walks

Aleksandar Bojchevski,Oleksandr Shchur,Daniel Zügner,Stephan Günnemann

We propose GraphGAN - the first implicit generative model for graphs that enables to mimic real-world networks.

We pose the problem of graph generation as learning the distribution of biased random walks over a single input graph.

Our model is based on a stochastic neural network that generates discrete output samples, and is trained using the Wasserstein GAN objective. GraphGAN enables us to generate sibling graphs, which have similar properties yet are not exact replicas of the original graph. Moreover, GraphGAN learns a semantic mapping from the latent input space to the generated graph's properties. We discover that sampling from certain regions of the latent space leads to varying properties of the output graphs, with smooth transitions between them. Strong generalization properties of GraphGAN are highlighted by its competitive performance in link prediction as well as promising results on node classification, even though not specifically trained for these tasks.

****

Pointing Out SQL Queries From Text

Chenglong Wang,Marc Brockschmidt,Rishabh Singh

The digitization of data has resulted in making datasets available to millions of users in the form of relational databases and spreadsheet tables. However, a majority of these users come from diverse backgrounds and lack the programming expertise to query and analyze such tables. We present a system that allows for querying data tables using natural language questions, where the system translates the question into an executable SQL query. We use a deep sequence to sequence m

odel in wich the decoder uses a simple type system of SQL expressions to structure the output prediction. Based on the type, the decoder either copies an output token from the input question using an attention-based copying mechanism or generates it from a fixed vocabulary. We also introduce a value-based loss function that transforms a distribution over locations to copy from into a distribution over the set of input tokens to improve training of our model. We evaluate our model on the recently released WikiSQL dataset and show that our model trained using only supervised learning significantly outperforms the current state-of-the-art Seq2SQL model that uses reinforcement learning.
****
Loss Functions for Multiset Prediction
Sean Welleck,Zixin Yao,Yu Gai,Jialin Mao,Zheng Zhang,Kyunghyun Cho
We study the problem of multiset prediction. The goal of multiset prediction is to train a predictor that maps an input to a multiset consisting of multiple items. Unlike existing problems in supervised learning, such as classification, ranking and sequence generation, there is no known order among items in a target multiset, and each item in the multiset may appear more than once, making this problem extremely challenging. In this paper, we propose a novel multiset loss function by viewing this problem from the perspective of sequential decision making. The proposed multiset loss function is empirically evaluated on two families of datasets, one synthetic and the other real, with varying levels of difficulty, against various baseline loss functions including reinforcement learning, sequence, and aggregated distribution matching loss functions. The experiments reveal the effectiveness of the proposed loss function over the others.
****
Mitigating Adversarial Effects Through Randomization
Cihang Xie,Jianyu Wang,Zhishuai Zhang,Zhou Ren,Alan Yuille
Convolutional neural networks have demonstrated high accuracy on various tasks in recent years. However, they are extremely vulnerable to adversarial examples. For example, imperceptible perturbations added to clean images can cause convolutional neural networks to fail. In this paper, we propose to utilize randomization at inference time to mitigate adversarial effects. Specifically, we use two randomization operations: random resizing, which resizes the input images to a random size, and random padding, which pads zeros around the input images in a random manner. Extensive experiments demonstrate that the proposed randomization method is very effective at defending against both single-step and iterative attacks. Our method provides the following advantages: 1) no additional training or fine-tuning, 2) very few additional computations, 3) compatible with other adversarial defense methods. By combining the proposed randomization method with an adversarially trained model, it achieves a normalized score of 0.924 (ranked No.2 among 107 defense teams) in the NIPS 2017 adversarial examples defense challenge, which is far better than using adversarial training alone with a normalized score of 0.773 (ranked No.56). The code is public available at https://github.com/cihangxie/NIPS2017_adv_challenge_defense.
****
Multi-Task Learning by Deep Collaboration and Application in Facial Landmark Detection
Ludovic Trottier,Philippe Giguère,Brahim Chaib-draa
Convolutional neural networks (CNN) have become the most successful and popular approach in many vision-related domains. While CNNs are particularly well-suited for capturing a proper hierarchy of concepts from real-world images, they are limited to domains where data is abundant. Recent attempts have looked into mitigating this data scarcity problem by casting their original single-task problem into a new multi-task learning (MTL) problem. The main goal of this inductive transfer mechanism is to leverage domain-specific information from related tasks, in order to improve generalization on the main task. While recent results in the deep learning (DL) community have shown the promising potential of training task-specific CNNs in a soft parameter sharing framework, integrating the recent DL advances for improving knowledge sharing is still an open problem. In this paper, we propose the Deep Collaboration Network (DCNet), a novel approach for connec

ting task-specific CNNs in a MTL framework. We define connectivity in terms of two distinct non-linear transformation blocks. One aggregates task-specific features into global features, while the other merges back the global features with each task-specific network. Based on the observation that task relevance depends on depth, our transformation blocks use skip connections as suggested by residual network approaches, to more easily deactivate unrelated task-dependent features. To validate our approach, we employed facial landmark detection (FLD) datasets as they are readily amenable to MTL, given the number of tasks they include. Experimental results show that we can achieve up to 24.31% relative improvement in landmark failure rate over other state-of-the-art MTL approaches. We finally perform an ablation study showing that our approach effectively allows knowledge sharing, by leveraging domain-specific features at particular depths from tasks that we know are related.
****

Reinforcement and Imitation Learning for Diverse Visuomotor Skills
Yuke Zhu,Ziyu Wang,Josh Merel,Andrei Rusu,Tom Erez,Serkan Cabi,Saran Tunyasuvunakool,János Kramár,Raia Hadsell,Nando de Freitas,Nicolas Heess
We propose a general deep reinforcement learning method and apply it to robot manipulation tasks. Our approach leverages demonstration data to assist a reinforcement learning agent in learning to solve a wide range of tasks, mainly previously unsolved. We train visuomotor policies end-to-end to learn a direct mapping from RGB camera inputs to joint velocities. Our experiments indicate that our reinforcement and imitation approach can solve contact-rich robot manipulation tasks that neither the state-of-the-art reinforcement nor imitation learning method can solve alone. We also illustrate that these policies achieved zero-shot sim2real transfer by training with large visual and dynamics variations.
****

DropMax: Adaptive Stochastic Softmax
Hae Beom Lee,Juho Lee,Eunho Yang,Sung Ju Hwang
We propose DropMax, a stochastic version of softmax classifier which at each iteration drops non-target classes with some probability, for each instance. Specifically, we overlay binary masking variables over class output probabilities, which are learned based on the input via regularized variational inference. This stochastic regularization has an effect of building an ensemble classifier out of combinatorial number of classifiers with different decision boundaries. Moreover, the learning of dropout probabilities for non-target classes on each instance allows the classifier to focus more on classification against the most confusing classes. We validate our model on multiple public datasets for classification, on which it obtains improved accuracy over regular softmax classifier and other baselines. Further analysis of the learned dropout masks shows that our model indeed selects confusing classes more often when it performs classification.
****

Composable Planning with Attributes
Amy Zhang,Adam Lerer,Sainbayar Sukhbaatar,Rob Fergus,Arthur Szlam
The tasks that an agent will need to solve often aren’t known during training. However, if the agent knows which properties of the environment we consider im- portant, then after learning how its actions affect those properties the agent may be able to use this knowledge to solve complex tasks without training specifi- cally for them. Towards this end, we consider a setup in which an environment is augmented with a set of user defined attributes that parameterize the features of interest. We propose a model that learns a policy for transitioning between “nearby” sets of attributes, and maintains a graph of possible transitions. Given a task at test time that can be expressed in terms of a target set of attributes, and a current state, our model infers the attributes of the current state and searches over paths through attribute space to get a high level plan, and then uses its low level policy to execute the plan. We show in grid-world games and 3D block stacking that our model is able to generalize to longer, more complex tasks at test time even when it only sees short, simple tasks at train time.

****

Attention-based Graph Neural Network for Semi-supervised Learning
Kiran K. Thekumparampil,Sewoong Oh,Chong Wang,Li-Jia Li
Recently popularized graph neural networks achieve the state-of-the-art accuracy on a number of standard benchmark datasets for graph-based semi-supervised learning, improving significantly over existing approaches. These architectures alternate between a propagation layer that aggregates the hidden states of the local neighborhood and a fully-connected layer. Perhaps surprisingly, we show that a linear model, that removes all the intermediate fully-connected layers, is still able to achieve a performance comparable to the state-of-the-art models. This significantly reduces the number of parameters, which is critical for semi-supervised learning where number of labeled examples are small. This in turn allows a room for designing more innovative propagation layers. Based on this insight, we propose a novel graph neural network that removes all the intermediate fully-connected layers, and replaces the propagation layers with attention mechanisms that respect the structure of the graph. The attention mechanism allows us to learn a dynamic and adaptive local summary of the neighborhood to achieve more accurate predictions. In a number of experiments on benchmark citation networks datasets, we demonstrate that our approach outperforms competing methods. By examining the attention weights among neighbors, we show that our model provides some interesting insights on how neighbors influence each other.
****
Learning an Embedding Space for Transferable Robot Skills
Karol Hausman,Jost Tobias Springenberg,Ziyu Wang,Nicolas Heess,Martin Riedmiller
We present a method for reinforcement learning of closely related skills that are parameterized via a skill embedding space. We learn such skills by taking advantage of latent variables and exploiting a connection between reinforcement learning and variational inference. The main contribution of our work is an entropy-regularized policy gradient formulation for hierarchical policies, and an associated, data-efficient and robust off-policy gradient algorithm based on stochastic value gradients. We demonstrate the effectiveness of our method on several simulated robotic manipulation tasks. We find that our method allows for discovery of multiple solutions and is capable of learning the minimum number of distinct skills that are necessary to solve a given set of tasks. In addition, our results indicate that the hereby proposed technique can interpolate and/or sequence previously learned skills in order to accomplish more complex tasks, even in the presence of sparse rewards.

****
Learning Representations and Generative Models for 3D Point Clouds
Panos Achlioptas,Olga Diamanti,Ioannis Mitliagkas,Leonidas Guibas
Three-dimensional geometric data offer an excellent domain for studying representation learning and generative modeling. In this paper, we look at geometric data represented as point clouds. We introduce a deep autoencoder (AE) network with excellent reconstruction quality and generalization ability. The learned representations outperform the state of the art in 3D recognition tasks and enable basic shape editing applications via simple algebraic manipulations, such as semantic part editing, shape analogies and shape interpolation. We also perform a thorough study of different generative models including GANs operating on the raw point clouds, significantly improved GANs trained in the fixed latent space our AEs and, Gaussian mixture models (GMM). Interestingly, GMMs trained in the latent space of our AEs produce samples of the best fidelity and diversity.
To perform our quantitative evaluation of generative models, we propose simple measures of fidelity and diversity based on optimally matching between sets point clouds.
****
State Space LSTM Models with Particle MCMC Inference
Xun Zheng,Manzil Zaheer,Amr Ahmed,Yuan Wang,Eric P. Xing,Alex Smola
Long Short-Term Memory (LSTM) is one of the most powerful sequence models. Despite the strong performance, however, it lacks the nice interpretability as in state space models. In this paper, we present a way to combine the best of both wor

lds by introducing State Space LSTM (SSL), which generalizes the earlier work \cite{zaheer2017latent} of combining topic models with LSTM. However, unlike \cite{zaheer2017latent}, we do not make any factorization assumptions in our inference algorithm. We present an efficient sampler based on sequential Monte Carlo (SMC) method that draws from the joint posterior directly. Experimental results confirms the superiority and stability of this SMC inference algorithm on a variety of domains.
****
Theoretical properties of the global optimizer of two-layer Neural Network
Digvijay Boob,Guanghui Lan
In this paper, we study the problem of optimizing a two-layer artificial neural network that best fits a training dataset. We look at this problem in the setting where the number of parameters is greater than the number of sampled points. We show that for a wide class of differentiable activation functions (this class involves most nonlinear functions and excludes piecewise linear functions), we have that arbitrary first-order optimal solutions satisfy global optimality provided the hidden layer is non-singular. We essentially show that these non-singular hidden layer matrix satisfy a ``"good" property for these big class of activation functions. Techniques involved in proving this result inspire us to look at a new algorithmic, where in between two gradient step of hidden layer, we add a stochastic gradient descent (SGD) step of the output layer. In this new algorithmic framework, we extend our earlier result and show that for all finite iterations the hidden layer satisfies the``good" property mentioned earlier therefore partially explaining success of noisy gradient methods and addressing the issue of data independency of our earlier result. Both of these results are easily extended to hidden layers given by a flat matrix from that of a square matrix. Results are applicable even if network has more than one hidden layer provided all inner hidden layers are arbitrary, satisfy non-singularity, all activations are from the given class of differentiable functions and optimization is only with respect to the outermost hidden layer. Separately, we also study the smoothness properties of the objective function and show that it is actually Lipschitz smooth, i.e., its gradients do not change sharply. We use smoothness properties to guarantee asymptotic convergence of $O(1/\text{number of iterations})$ to a first-order optimal solution.
****
Learning to cluster in order to transfer across domains and tasks
Yen-Chang Hsu,Zhaoyang Lv,Zsolt Kira
This paper introduces a novel method to perform transfer learning across domains and tasks, formulating it as a problem of learning to cluster. The key insight is that, in addition to features, we can transfer similarity information and this is sufficient to learn a similarity function and clustering network to perform both domain adaptation and cross-task transfer learning. We begin by reducing categorical information to pairwise constraints, which only considers whether two instances belong to the same class or not (pairwise semantic similarity). This similarity is category-agnostic and can be learned from data in the source domain using a similarity network. We then present two novel approaches for performing transfer learning using this similarity function. First, for unsupervised domain adaptation, we design a new loss function to regularize classification with a constrained clustering loss, hence learning a clustering network with the transferred similarity metric generating the training inputs. Second, for cross-task learning (i.e., unsupervised clustering with unseen categories), we propose a framework to reconstruct and estimate the number of semantic clusters, again using the clustering network. Since the similarity network is noisy, the key is to use a robust clustering algorithm, and we show that our formulation is more robust than the alternative constrained and unconstrained clustering approaches. Using this method, we first show state of the art results for the challenging cross-task problem, applied on Omniglot and ImageNet. Our results show that we can reconstruct semantic clusters with high accuracy. We then evaluate the performance of cross-domain transfer using images from the Office-31 and SVHN-MNIST tasks and present top accuracy on both datasets. Our approach doesn't explicitly deal wi

th domain discrepancy. If we combine with a domain adaptation loss, it shows further improvement.
****
Spatially Transformed Adversarial Examples
Chaowei Xiao,Jun-Yan Zhu,Bo Li,Warren He,Mingyan Liu,Dawn Song
Recent studies show that widely used Deep neural networks (DNNs) are vulnerable to the carefully crafted adversarial examples.
Many advanced algorithms have been proposed to generate adversarial examples by leveraging the L_p distance for penalizing perturbations.
Different defense methods have also been explored to defend against such adversarial attacks.
While the effectiveness of L_p distance as a metric of perceptual quality remains an active research area, in this paper we will instead focus on a different type of perturbation, namely spatial transformation, as opposed to manipulating the pixel values directly as in prior works.
Perturbations generated through spatial transformation could result in large L_p distance measures, but our extensive experiments show that such spatially transformed adversarial examples are perceptually realistic and more difficult to defend against with existing defense systems. This potentially provides a new direction in adversarial example generation and the design of corresponding defenses.
We visualize the spatial transformation based perturbation for different examples and show that our technique
can produce realistic adversarial examples with smooth image deformation.
Finally, we visualize the attention of deep networks with different types of adversarial examples to better understand how these examples are interpreted.
****
Semi-supervised Outlier Detection using Generative And Adversary Framework
Jindong Gu,Matthias Schubert,Volker Tresp
In a conventional binary/multi-class classification task, the decision boundary is supported by data from two or more classes. However, in one-class classification task, only data from one class are available. To build an robust outlier detector using only data from a positive class, we propose a corrupted GAN(CorGAN), a deep convolutional Generative Adversary Network requiring no convergence during training. In the adversarial process of training CorGAN, the Generator is supposed to generate outlier samples for negative class, and the Discriminator as an one-class classifier is trained to distinguish data from training datasets (i.e. positive class) and generated data from the Generator (i.e. negative class). To improve the performance of the Discriminator (one-class classifier), we also propose a lot of techniques to improve the performance of the model. The proposed model outperforms the traditional method PCA + PSVM and the solution based on Autoencoder.
****
Graph Topological Features via GAN
Weiyi Liu,Hal Cooper,Min-Hwan Oh
Inspired by the success of generative adversarial networks (GANs) in image domains, we introduce a novel hierarchical architecture for learning characteristic topological features from a single arbitrary input graph via GANs. The hierarchical architecture consisting of multiple GANs preserves both local and global topological features, and automatically partitions the input graph into representative stages for feature learning. The stages facilitate reconstruction and can be used as indicators of the importance of the associated topological structures. Experiments show that our method produces subgraphs retaining a wide range of topological features, even in early reconstruction stages. This paper contains original research on combining the use of GANs and graph topological analysis.
****
Overcoming the vanishing gradient problem in plain recurrent networks
Yuhuang Hu,Adrian Huber,Shih-Chii Liu
Plain recurrent networks greatly suffer from the vanishing gradient problem while Gated Neural Networks (GNNs) such as Long-short Term Memory (LSTM) and Gated Recurrent Unit (GRU) deliver promising results in many sequence learning tasks th

rough sophisticated network designs. This paper shows how we can address this pr oblem in a plain recurrent network by analyzing the gating mechanisms in GNNs. W e propose a novel network called the Recurrent Identity Network (RIN) which allo ws a plain recurrent network to overcome the vanishing gradient problem while tr aining very deep models without the use of gates. We compare this model with IRN Ns and LSTMs on multiple sequence modeling benchmarks. The RINs demonstrate comp etitive performance and converge faster in all tasks. Notably, small RIN models produce 12%–67% higher accuracy on the Sequential and Permuted MNIST datasets an d reach state-of-the-art performance on the bAbI question answering dataset.
****

Fast and Accurate Inference with Adaptive Ensemble Prediction for Deep Networks
Hiroshi Inoue
Ensembling multiple predictions is a widely-used technique to improve the accura cy of various machine learning tasks. In image classification tasks, for example , averaging the predictions for multiple patches extracted from the input image significantly improves accuracy. Using multiple networks trained independently t o make predictions improves accuracy further. One obvious drawback of the ensemb ling technique is its higher execution cost during inference.% If we average 100 local predictions, the execution cost will be 100 times as high as the cost wit hout the ensemble. This higher cost limits the real-world use of ensembling. In this paper, we first describe our insights on relationship between the probabili ty of the prediction and the effect of ensembling with current deep neural netwo rks; ensembling does not help mispredictions for inputs predicted with a high pr obability, i.e. the output from the softmax. This finding motivates us to develo p a new technique called adaptive ensemble prediction, which achieves the benefi ts of ensembling with much smaller additional execution costs. Hence, we calcula te the confidence level of the prediction for each input from the probabilities of the local predictions during the ensembling computation. If the prediction fo r an input reaches a high enough probability on the basis of the confidence leve l, we stop ensembling for this input to avoid wasting computation power. We eval uated the adaptive ensembling by using various datasets and showed that it reduc es the computation cost significantly while achieving similar accuracy to the na ive ensembling. We also showed that our statistically rigorous confidence-level-based termination condition reduces the burden of the task-dependent parameter t uning compared to the naive termination based on the pre-defined threshold in ad dition to yielding a better accuracy with the same cost.

****

GANITE: Estimation of Individualized Treatment Effects using Generative Adversar ial Nets
Jinsung Yoon,James Jordon,Mihaela van der Schaar
Estimating individualized treatment effects (ITE) is a challenging task due to t he need for an individual's potential outcomes to be learned from biased data an d without having access to the counterfactuals. We propose a novel method for in ferring ITE based on the Generative Adversarial Nets (GANs) framework. Our metho d, termed Generative Adversarial Nets for inference of Individualized Treatment Effects (GANITE), is motivated by the possibility that we can capture the uncert ainty in the counterfactual distributions by attempting to learn them using a GA N. We generate proxies of the counterfactual outcomes using a counterfactual gen erator, G, and then pass these proxies to an ITE generator, I, in order to train it. By modeling both of these using the GAN framework, we are able to infer bas ed on the factual data, while still accounting for the unseen counterfactuals. W e test our method on three real-world datasets (with both binary and multiple tr eatments) and show that GANITE outperforms state-of-the-art methods.
****

Overcoming Catastrophic Interference using Conceptor-Aided Backpropagation
Xu He,Herbert Jaeger
Catastrophic interference has been a major roadblock in the research of continua l learning. Here we propose a variant of the back-propagation algorithm, "Concep tor-Aided Backprop" (CAB), in which gradients are shielded by conceptors against

 degradation of previously learned tasks. Conceptors have their origin in reserv
oir computing, where they have been previously shown to overcome catastrophic fo
rgetting. CAB extends these results to deep feedforward networks. On the disjoin
t and permuted MNIST tasks, CAB outperforms two other methods for coping with ca
tastrophic interference that have recently been proposed.
****
Inference Suboptimality in Variational Autoencoders
Chris Cremer,Xuechen Li,David Duvenaud
Amortized inference has led to efficient approximate inference for large dataset
s. The quality of posterior inference is largely determined by two factors: a) t
he ability of the variational distribution to model the true posterior and b) th
e capacity of the recognition network to generalize inference over all datapoint
s. We analyze approximate inference in variational autoencoders in terms of thes
e factors. We find that suboptimal inference is often due to amortizing inferenc
e rather than the limited complexity of the approximating distribution. We show
that this is due partly to the generator learning to accommodate the choice of a
pproximation. Furthermore, we show that the parameters used to increase the expr
essiveness of the approximation play a role in generalizing inference rather tha
n simply improving the complexity of the approximation.
****
Initialization matters: Orthogonal Predictive State Recurrent Neural Networks
Krzysztof Choromanski,Carlton Downey,Byron Boots
Learning to predict complex time-series data is a fundamental challenge in a ran
ge of disciplines including Machine Learning, Robotics, and Natural Language Pro
cessing. Predictive State Recurrent Neural Networks (PSRNNs) (Downey et al.) are
 a state-of-the-art approach for modeling time-series data which combine the ben
efits of probabilistic filters and Recurrent Neural Networks into a single model
. PSRNNs leverage the concept of Hilbert Space Embeddings of distributions (Smol
a et al.) to embed predictive states into a Reproducing Kernel Hilbert Space, th
en estimate, predict, and update these embedded states using Kernel Bayes Rule.
Practical implementations of PSRNNs are made possible by the machinery of Random
 Features, where input features are mapped into a new space where dot products a
pproximate the kernel well. Unfortunately PSRNNs often require a large number of
 RFs to obtain good results, resulting in large models which are slow to execute
 and slow to train. Orthogonal Random Features (ORFs) (Choromanski et al.) is an
 improvement on RFs which has been shown to decrease the number of RFs required
for pointwise kernel approximation. Unfortunately, it is not clear that ORFs can
 be applied to PSRNNs, as PSRNNs rely on Kernel Ridge Regression as a core compo
nent of their learning algorithm, and the theoretical guarantees of ORF do not a
pply in this setting. In this paper, we extend the theory of ORFs to Kernel Ridg
e Regression and show that ORFs can be used to obtain Orthogonal PSRNNs (OPSRNNs
), which are smaller and faster than PSRNNs. In particular, we show that OPSRNN
models clearly outperform LSTMs and furthermore, can achieve accuracy similar to
 PSRNNs with an order of magnitude smaller number of features needed.
****
Discrete Sequential Prediction of Continuous Actions for Deep RL
Luke Metz,Julian Ibarz,Navdeep Jaitly,James Davidson
It has long been assumed that high dimensional continuous control problems canno
t be solved effectively by discretizing individual dimensions of the action spac
e due to the exponentially large number of bins over which policies would have t
o be learned. In this paper, we draw inspiration from the recent success of sequ
ence-to-sequence models for structured prediction problems to develop policies o
ver discretized spaces. Central to this method is the realization that complex f
unctions over high dimensional spaces can be modeled by neural networks that pre
dict one dimension at a time. Specifically, we show how Q-values and policies ov
er continuous spaces can be modeled using a next step prediction model over disc
retized dimensions. With this parameterization, it is possible to both leverage
the compositional structure of action spaces during learning, as well as compute
 maxima over action spaces (approximately). On a simple example task we demonstr
ate empirically that our method can perform global search, which effectively get

s around the local optimization issues that plague DDPG. We apply the technique to off-policy (Q-learning) methods and show that our method can achieve the state-of-the-art for off-policy methods on several continuous control tasks.
****

Convolutional Normalizing Flows
Guoqing Zheng,Yiming Yang,Jaime Carbonell
Bayesian posterior inference is prevalent in various machine learning problems. Variational inference provides one way to approximate the posterior distribution, however its expressive power is limited and so is the accuracy of resulting approximation. Recently, there has a trend of using neural networks to approximate the variational posterior distribution due to the flexibility of neural network architecture. One way to construct flexible variational distribution is to warp a simple density into a complex by normalizing flows, where the resulting density can be analytically evaluated. However, there is a trade-off between the flexibility of normalizing flow and computation cost for efficient transformation. In this paper, we propose a simple yet effective architecture of normalizing flows, ConvFlow, based on convolution over the dimensions of random input vector. Experiments on synthetic and real world posterior inference problems demonstrate the effectiveness and efficiency of the proposed method.
****

Neural Tree Transducers for Tree to Tree Learning
João Sedoc,Dean Foster,Lyle Ungar
We introduce a novel approach to tree-to-tree learning, the neural tree transducer (NTT), a top-down depth first context-sensitive tree decoder, which is paired with recursive neural encoders. Our method works purely on tree-to-tree manipulations rather than sequence-to-tree or tree-to-sequence and is able to encode and decode multiple depth trees. We compare our method to sequence-to-sequence models applied to serializations of the trees and show that our method outperforms previous methods for tree-to-tree transduction.
****

Forward Modeling for Partial Observation Strategy Games - A StarCraft Defogger
Gabriel Synnaeve,Zeming Lin,Jonas Gehring,Vasil Khalidov,Nicolas Carion,Nicolas Usunier
This paper we present a defogger, a model that learns to predict future hidden information from partial observations. We formulate this model in the context of forward modeling and leverage spatial and sequential constraints and correlations via convolutional neural networks and long short-term memory networks, respectively. We evaluate our approach on a large dataset of human games of StarCraft: Brood War, a real-time strategy video game. Our models consistently beat strong rule-based baselines and qualitatively produce sensible future game states.
****

Explaining the Mistakes of Neural Networks with Latent Sympathetic Examples
Riaan Zoetmulder,Efstratios Gavves,Peter O'Connor
Neural networks make mistakes. The reason why a mistake is made often remains a mystery. As such neural networks often are considered a black box. It would be useful to have a method that can give an explanation that is intuitive to a user as to why an image is misclassified. In this paper we develop a method for explaining the mistakes of a classifier model by visually showing what must be added to an image such that it is correctly classified. Our work combines the fields of adversarial examples, generative modeling and a correction technique based on difference target propagation to create an technique that creates explanations of why an image is misclassified. In this paper we explain our method and demonstrate it on MNIST and CelebA. This approach could aid in demystifying neural networks for a user.

****

Fidelity-Weighted Learning
Mostafa Dehghani,Arash Mehrjou,Stephan Gouws,Jaap Kamps,Bernhard Schölkopf
Training deep neural networks requires many training samples, but in practice training labels are expensive to obtain and may be of varying quality, as some may

 be from trusted expert labelers while others might be from heuristics or other sources of weak supervision such as crowd-sourcing. This creates a fundamental q uality- versus-quantity trade-off in the learning process. Do we learn from the small amount of high-quality data or the potentially large amount of weakly-labe led data? We argue that if the learner could somehow know and take the label-qua lity into account when learning the data representation, we could get the best o f both worlds. To this end, we propose “fidelity-weighted learning” (FWL), a sem i-supervised student- teacher approach for training deep neural networks using w eakly-labeled data. FWL modulates the parameter updates to a student network (tr ained on the task we care about) on a per-sample basis according to the posterio r confidence of its label-quality estimated by a teacher (who has access to the high-quality labels). Both student and teacher are learned from the data. We eva luate FWL on two tasks in information retrieval and natural language processing where we outperform state-of-the-art alternative semi-supervised methods, indica ting that our approach makes better use of strong and weak labels, and leads to better task-dependent data representations.
****

Parametric Manifold Learning Via Sparse Multidimensional Scaling
Gautam Pai,Ronen Talmon,Ron Kimmel
We propose a metric-learning framework for computing distance-preserving maps th at generate low-dimensional embeddings for a certain class of manifolds. We empl oy Siamese networks to solve the problem of least squares multidimensional scali ng for generating mappings that preserve geodesic distances on the manifold. In contrast to previous parametric manifold learning methods we show a substantial reduction in training effort enabled by the computation of geodesic distances in a farthest point sampling strategy. Additionally, the use of a network to model the distance-preserving map reduces the complexity of the multidimensional scal ing problem and leads to an improved non-local generalization of the manifold co mpared to analogous non-parametric counterparts. We demonstrate our claims on po int-cloud data and on image manifolds and show a numerical analysis of our techn ique to facilitate a greater understanding of the representational power of neur al networks in modeling manifold data.
****

Adversarial Spheres
Justin Gilmer,Luke Metz,Fartash Faghri,Sam Schoenholz,Maithra Raghu,Martin Watte nberg,Ian Goodfellow
    State of the art computer vision models have been shown to be vulnerable to small adversarial perturbations of the input. In other words, most images in the data distribution are both correctly classified by the model and are very close to a visually similar misclassified image. Despite substantial research interes t, the cause of the phenomenon is still poorly understood and remains unsolved. We hypothesize that this counter intuitive behavior is a naturally occurring res ult of the high dimensional geometry of the data manifold. As a first step towar ds exploring this hypothesis, we study a simple synthetic dataset of classifying between two concentric high dimensional spheres. For this dataset we show a fun damental tradeoff between the amount of test error and the average distance to n earest error. In particular, we prove that any model which misclassifies a small constant fraction of a sphere will be vulnerable to adversarial perturbations o f size $O(1/\sqrt{d})$. Surprisingly, when we train several different architectu res on this dataset, all of their error sets naturally approach this theoretical bound. As a result of the theory, the vulnerability of neural networks to small adversarial perturbations is a logical consequence of the amount of test error observed. We hope that our theoretical analysis of this very simple case will po int the way forward to explore how the geometry of complex real-world data sets leads to adversarial examples.
****

Towards Effective GANs for Data Distributions with Diverse Modes
Sanchit Agrawal,Gurneet Singh,Mitesh Khapra
Generative Adversarial Networks (GANs), when trained on large datasets with dive rse modes, are known to produce conflated images which do not distinctly belong

to any of the modes. We hypothesize that this problem occurs due to the interaction between two facts: (1) For datasets with large variety, it is likely that the modes lie on separate manifolds. (2) The generator (G) is formulated as a continuous function, and the input noise is derived from a connected set, due to which G's output is a connected set. If G covers all modes, then there must be some portion of G's output which connects them. This corresponds to undesirable, conflated images. We develop theoretical arguments to support these intuitions. We propose a novel method to break the second assumption via learnable discontinuities in the latent noise space. Equivalently, it can be viewed as training several generators, thus creating discontinuities in the G function. We also augment the GAN formulation with a classifier C that predicts which noise partition/generator produced the output images, encouraging diversity between each partition/generator. We experiment on MNIST, celebA, STL-10, and a difficult dataset with clearly distinct modes, and show that the noise partitions correspond to different modes of the data distribution, and produce images of superior quality.
****
Learning General Purpose Distributed Sentence Representations via Large Scale Multi-task Learning
Sandeep Subramanian,Adam Trischler,Yoshua Bengio,Christopher J Pal
A lot of the recent success in natural language processing (NLP) has been driven by distributed vector representations of words trained on large amounts of text in an unsupervised manner. These representations are typically used as general purpose features for words across a range of NLP problems. However, extending this success to learning representations of sequences of words, such as sentences, remains an open problem. Recent work has explored unsupervised as well as supervised learning techniques with different training objectives to learn general purpose fixed-length sentence representations. In this work, we present a simple, effective multi-task learning framework for sentence representations that combines the inductive biases of diverse training objectives in a single model.
We train this model on several data sources with multiple training objectives on over 100 million sentences. Extensive experiments demonstrate that sharing a single recurrent sentence encoder across weakly related tasks leads to consistent improvements over previous methods. We present substantial improvements in the context of transfer learning and low-resource settings using our learned general-purpose representations.
****
Understanding Deep Neural Networks with Rectified Linear Units
Raman Arora,Amitabh Basu,Poorya Mianjy,Anirbit Mukherjee
In this paper we investigate the family of functions representable by deep neural networks (DNN) with rectified linear units (ReLU). We give an algorithm to train a ReLU DNN with one hidden layer to {\em global optimality} with runtime polynomial in the data size albeit exponential in the input dimension. Further, we improve on the known lower bounds on size (from exponential to super exponential) for approximating a ReLU deep net function by a shallower ReLU net. Our gap theorems hold for smoothly parametrized families of ``hard'' functions, contrary to countable, discrete families known in the literature. An example consequence of our gap theorems is the following: for every natural number k there exists a function representable by a ReLU DNN with k^2 hidden layers and total size k^3, such that any ReLU DNN with at most k hidden layers will require at least $\frac12k^{k+1}-1$ total nodes. Finally, for the family of $\R^n\to \R$ DNNs with ReLU activations, we show a new lowerbound on the number of affine pieces, which is larger than previous constructions in certain regimes of the network architecture and most distinctively our lowerbound is demonstrated by an explicit construction of a \emph{smoothly parameterized} family of functions attaining this scaling. Our construction utilizes the theory of zonotopes from polyhedral theory.
****
Enhancing the Transferability of Adversarial Examples with Noise Reduced Gradient
Lei Wu,Zhanxing Zhu,Cheng Tai,Weinan E

Deep neural networks provide state-of-the-art performance for many applications of interest. Unfortunately they are known to be vulnerable to adversarial examples, formed by applying small but malicious perturbations to the original inputs. Moreover, the perturbations can transfer across models: adversarial examples generated for a specific model will often mislead other unseen models. Consequently the adversary can leverage it to attack against the deployed black-box systems.

In this work, we demonstrate that the adversarial perturbation can be decomposed into two components: model-specific and data-dependent one, and it is the latter that mainly contributes to the transferability. Motivated by this understanding, we propose to craft adversarial examples by utilizing the noise reduced gradient (NRG) which approximates the data-dependent component. Experiments on various classification models trained on ImageNet demonstrates that the new approach enhances the transferability dramatically. We also find that low-capacity models have more powerful attack capability than high-capacity counterparts, under the condition that they have comparable test performance. These insights give rise to a principled manner to construct adversarial examples with high success rates and could potentially provide us guidance for designing effective defense approaches against black-box attacks.

****

Jiffy: A Convolutional Approach to Learning Time Series Similarity

Divya Shanmugam,Davis Blalock,John Guttag

Computing distances between examples is at the core of many learning algorithms for time series. Consequently, a great deal of work has gone into designing effective time series distance measures. We present Jiffy, a simple and scalable distance metric for multivariate time series. Our approach is to reframe the task as a representation learning problem---rather than design an elaborate distance function, we use a CNN to learn an embedding such that the Euclidean distance is effective. By aggressively max-pooling and downsampling, we are able to construct this embedding using a highly compact neural network. Experiments on a diverse set of multivariate time series datasets show that our approach consistently outperforms existing methods.

****

Improving GANs Using Optimal Transport

Tim Salimans,Han Zhang,Alec Radford,Dimitris Metaxas

We present Optimal Transport GAN (OT-GAN), a variant of generative adversarial nets minimizing a new metric measuring the distance between the generator distribution and the data distribution. This metric, which we call mini-batch energy distance, combines optimal transport in primal form with an energy distance defined in an adversarially learned feature space, resulting in a highly discriminative distance function with unbiased mini-batch gradients. Experimentally we show OT-GAN to be highly stable when trained with large mini-batches, and we present state-of-the-art results on several popular benchmark problems for image generation.

****

Spectral Normalization for Generative Adversarial Networks

Takeru Miyato,Toshiki Kataoka,Masanori Koyama,Yuichi Yoshida

One of the challenges in the study of generative adversarial networks is the instability of its training.

In this paper, we propose a novel weight normalization technique called spectral normalization to stabilize the training of the discriminator.

Our new normalization technique is computationally light and easy to incorporate into existing implementations.

We tested the efficacy of spectral normalization on CIFAR10, STL-10, and ILSVRC2012 dataset, and we experimentally confirmed that spectrally normalized GANs (SN-GANs) is capable of generating images of better or equal quality relative to the previous training stabilization techniques.

****

Building effective deep neural networks one feature at a time

Martin Mundt,Tobias Weis,Kishore Konda,Visvanathan Ramesh

Successful training of convolutional neural networks is often associated with su
ffi-
ciently deep architectures composed of high amounts of features. These networks
typically rely on a variety of regularization and pruning techniques to converge
to less redundant states. We introduce a novel bottom-up approach to expand
representations in fixed-depth architectures. These architectures start from jus
t a
single feature per layer and greedily increase width of individual layers to att
ain
effective representational capacities needed for a specific task. While network
growth can rely on a family of metrics, we propose a computationally efficient
version based on feature time evolution and demonstrate its potency in determin-
ing feature importance and a networks' effective capacity. We demonstrate how
automatically expanded architectures converge to similar topologies that benefit
from lesser amount of parameters or improved accuracy and exhibit systematic
correspondence in representational complexity with the specified task. In contra
st
to conventional design patterns with a typical monotonic increase in the amount
of
features with increased depth, we observe that CNNs perform better when there is
more learnable parameters in intermediate, with falloffs to earlier and later la
yers.
****

Routing Networks: Adaptive Selection of Non-Linear Functions for Multi-Task Lear
ning

Clemens Rosenbaum,Tim Klinger,Matthew Riemer

Multi-task learning (MTL) with neural networks leverages commonalities in tasks
to improve performance, but often suffers from task interference which reduces t
he benefits of transfer. To address this issue we introduce the routing network
paradigm, a novel neural network and training algorithm. A routing network is a
kind of self-organizing neural network consisting of two components: a router an
d a set of one or more function blocks. A function block may be any neural netwo
rk – for example a fully-connected or a convolutional layer. Given an input the
router makes a routing decision, choosing a function block to apply and passing
the output back to the router recursively, terminating when a fixed recursion de
pth is reached. In this way the routing network dynamically composes different f
unction blocks for each input. We employ a collaborative multi-agent reinforceme
nt learning (MARL) approach to jointly train the router and function blocks. We
evaluate our model against cross-stitch networks and shared-layer baselines on m
ulti-task settings of the MNIST, mini-imagenet, and CIFAR-100 datasets. Our expe
riments demonstrate a significant improvement in accuracy, with sharper converge
nce. In addition, routing networks have nearly constant per-task training cost w
hile cross-stitch networks scale linearly with the number of tasks. On CIFAR100
(20 tasks) we obtain cross-stitch performance levels with an 85% average reducti
on in training time.

****

On Unifying Deep Generative Models

Zhiting Hu,Zichao Yang,Ruslan Salakhutdinov,Eric P. Xing

Deep generative models have achieved impressive success in recent years. Generat
ive Adversarial Networks (GANs) and Variational Autoencoders (VAEs), as powerful
 frameworks for deep generative model learning, have largely been considered as
two distinct paradigms and received extensive independent studies respectively.
This paper aims to establish formal connections between GANs and VAEs through a
new formulation of them. We interpret sample generation in GANs as performing po
sterior inference, and show that GANs and VAEs involve minimizing KL divergences
 of respective posterior and inference distributions with opposite directions, e
xtending the two learning phases of classic wake-sleep algorithm, respectively.
The unified view provides a powerful tool to analyze a diverse set of existing m
odel variants, and enables to transfer techniques across research lines in a pri

ncipled way. For example, we apply the importance weighting method in VAE litera
tures for improved GAN learning, and enhance VAEs with an adversarial mechanism
that leverages generated samples. Experiments show generality and effectiveness
of the transfered techniques.
****
Learning to Represent Programs with Graphs
Miltiadis Allamanis,Marc Brockschmidt,Mahmoud Khademi
Learning tasks on source code (i.e., formal languages) have been considered rece
ntly, but most work has tried to transfer natural language methods and does not
capitalize on the unique opportunities offered by code's known syntax. For examp
le, long-range dependencies induced by using the same variable or function in di
stant locations are often not considered. We propose to use graphs to represent
both the syntactic and semantic structure of code and use graph-based deep learn
ing methods to learn to reason over program structures.

In this work, we present how to construct graphs from source code and how to sca
le Gated Graph Neural Networks training to such large graphs. We evaluate our me
thod on two tasks: VarNaming, in which a network attempts to predict the name of
 a variable given its usage, and VarMisuse, in which the network learns to reaso
n about selecting the correct variable that should be used at a given program lo
cation. Our comparison to methods that use less structured program representatio
ns shows the advantages of modeling known structure, and suggests that our model
s learn to infer meaningful names and to solve the VarMisuse task in many cases.
 Additionally, our testing showed that VarMisuse identifies a number of bugs in
mature open-source projects.
****
Improving the Improved Training of Wasserstein GANs: A Consistency Term and Its
Dual Effect
Xiang Wei,Boqing Gong,Zixia Liu,Wei Lu,Liqiang Wang
Despite being impactful on a variety of problems and applications, the generati
ve adversarial nets (GANs) are remarkably difficult to train. This issue is form
ally analyzed by \cite{arjovsky2017towards}, who also propose an alternative dir
ection to avoid the caveats in the minmax two-player training of GANs. The corre
sponding algorithm, namely, Wasserstein GAN (WGAN) hinges on the 1-Lipschitz con
tinuity of the discriminators. In this paper, we propose a novel approach for en
forcing the Lipschitz continuity in the training procedure of WGANs. Our approac
h seamlessly connects WGAN with one of the recent semi-supervised learning appro
aches. As a result, it gives rise to not only better photo-realistic samples t
han the previous methods but also state-of-the-art semi-supervised learning res
ults. In particular, to the best of our knowledge, our approach gives rise to t
he inception score of more than 5.0 with only 1,000 CIFAR10 images and is the fi
rst that exceeds the accuracy of 90\% the CIFAR10 datasets using only 4,000 labe
led images.

****
Unsupervised Neural Machine Translation
Mikel Artetxe,Gorka Labaka,Eneko Agirre,Kyunghyun Cho
In spite of the recent success of neural machine translation (NMT) in standard b
enchmarks, the lack of large parallel corpora poses a major practical problem fo
r many language pairs. There have been several proposals to alleviate this issue
 with, for instance, triangulation and semi-supervised learning techniques, but
they still require a strong cross-lingual signal. In this work, we completely re
move the need of parallel data and propose a novel method to train an NMT system
 in a completely unsupervised manner, relying on nothing but monolingual corpora
. Our model builds upon the recent work on unsupervised embedding mappings, and
consists of a slightly modified attentional encoder-decoder model that can be tr
ained on monolingual corpora alone using a combination of denoising and backtran
slation. Despite the simplicity of the approach, our system obtains 15.56 and 10
.21 BLEU points in WMT 2014 French-to-English and German-to-English translation.
 The model can also profit from small parallel corpora, and attains 21.81 and 15

.24 points when combined with 100,000 parallel sentences, respectively. Our implementation is released as an open source project.
****
Learning Independent Causal Mechanisms
Giambattista Parascandolo,Mateo Rojas Carulla,Niki Kilbertus,Bernhard Schoelkopf
Independent causal mechanisms are a central concept in the study of causality with implications for machine learning tasks. In this work we develop an algorithm to recover a set of (inverse) independent mechanisms relating a distribution transformed by the mechanisms to a reference distribution. The approach is fully unsupervised and based on a set of experts that compete for data to specialize and extract the mechanisms. We test and analyze the proposed method on a series of experiments based on image transformations. Each expert successfully maps a subset of the transformed data to the original domain, and the learned mechanisms generalize to other domains. We discuss implications for domain transfer and links to recent trends in generative modeling.
****
Deep Gaussian Embedding of Graphs: Unsupervised Inductive Learning via Ranking
Aleksandar Bojchevski,Stephan Günnemann
Methods that learn representations of nodes in a graph play a critical role in network analysis since they enable many downstream learning tasks. We propose Graph2Gauss - an approach that can efficiently learn versatile node embeddings on large scale (attributed) graphs that show strong performance on tasks such as link prediction and node classification. Unlike most approaches that represent nodes as point vectors in a low-dimensional continuous space, we embed each node as a Gaussian distribution, allowing us to capture uncertainty about the representation. Furthermore, we propose an unsupervised method that handles inductive learning scenarios and is applicable to different types of graphs: plain/attributed, directed/undirected. By leveraging both the network structure and the associated node attributes, we are able to generalize to unseen nodes without additional training. To learn the embeddings we adopt a personalized ranking formulation w.r.t. the node distances that exploits the natural ordering of the nodes imposed by the network structure. Experiments on real world networks demonstrate the high performance of our approach, outperforming state-of-the-art network embedding methods on several different tasks. Additionally, we demonstrate the benefits of modeling uncertainty - by analyzing it we can estimate neighborhood diversity and detect the intrinsic latent dimensionality of a graph.
****
Parametrized Hierarchical Procedures for Neural Programming
Roy Fox,Richard Shin,Sanjay Krishnan,Ken Goldberg,Dawn Song,Ion Stoica
Neural programs are highly accurate and structured policies that perform algorithmic tasks by controlling the behavior of a computation mechanism. Despite the potential to increase the interpretability and the compositionality of the behavior of artificial agents, it remains difficult to learn from demonstrations neural networks that represent computer programs. The main challenges that set algorithmic domains apart from other imitation learning domains are the need for high accuracy, the involvement of specific structures of data, and the extremely limited observability. To address these challenges, we propose to model programs as Parametrized Hierarchical Procedures (PHPs). A PHP is a sequence of conditional operations, using a program counter along with the observation to select between taking an elementary action, invoking another PHP as a sub-procedure, and returning to the caller. We develop an algorithm for training PHPs from a set of supervisor demonstrations, only some of which are annotated with the internal call structure, and apply it to efficient level-wise training of multi-level PHPs. We show in two benchmarks, NanoCraft and long-hand addition, that PHPs can learn neural programs more accurately from smaller amounts of both annotated and unannotated demonstrations.
****
Training and Inference with Integers in Deep Neural Networks
Shuang Wu,Guoqi Li,Feng Chen,Luping Shi

Researches on deep neural networks with discrete parameters and their deployment
 in embedded systems have been active and promising topics. Although previous wo
rks have successfully reduced precision in inference, transferring both training
 and inference processes to low-bitwidth integers has not been demonstrated simu
ltaneously. In this work, we develop a new method termed as ``"WAGE" to discreti
ze both training and inference, where weights (W), activations (A), gradients (G
) and errors (E) among layers are shifted and linearly constrained to low-bitwid
th integers. To perform pure discrete dataflow for fixed-point devices, we furth
er replace batch normalization by a constant scaling layer and simplify other co
mponents that are arduous for integer implementation. Improved accuracies can be
 obtained on multiple datasets, which indicates that WAGE somehow acts as a type
 of regularization. Empirically, we demonstrate the potential to deploy training
 in hardware systems such as integer-based deep learning accelerators and neurom
orphic chips with comparable accuracy and higher energy efficiency, which is cru
cial to future AI applications in variable scenarios with transfer and continual
 learning demands.
****
Eigenoption Discovery through the Deep Successor Representation
Marlos C. Machado,Clemens Rosenbaum,Xiaoxiao Guo,Miao Liu,Gerald Tesauro,Murray
Campbell
Options in reinforcement learning allow agents to hierarchically decompose a tas
k into subtasks, having the potential to speed up learning and planning. However
, autonomously learning effective sets of options is still a major challenge in
the field. In this paper we focus on the recently introduced idea of using repre
sentation learning methods to guide the option discovery process. Specifically,
we look at eigenoptions, options obtained from representations that encode diffu
sive information flow in the environment. We extend the existing algorithms for
eigenoption discovery to settings with stochastic transitions and in which handc
rafted features are not available. We propose an algorithm that discovers eigen
options while learning non-linear state representations from raw pixels. It expl
oits recent successes in the deep reinforcement learning literature and the equi
valence between proto-value functions and the successor representation. We use t
raditional tabular domains to provide intuition about our approach and Atari 260
0 games to demonstrate its potential.
****
Weighted Transformer Network for Machine Translation
Karim Ahmed,Nitish Shirish Keskar,Richard Socher
State-of-the-art results on neural machine translation often use attentional seq
uence-to-sequence models with some form of convolution or recursion. Vaswani et.
 al. (2017) propose a new architecture that avoids recurrence and convolution co
mpletely. Instead, it uses only self-attention and feed-forward layers. While th
e proposed architecture achieves state-of-the-art results on several machine tra
nslation tasks, it requires a large number of parameters and training iterations
 to converge. We propose Weighted Transformer, a Transformer with modified atten
tion layers, that not only outperforms the baseline network in BLEU score but al
so converges 15-40% faster. Specifically, we replace the multi-head attention by
 multiple self-attention branches that the model learns to combine during the tr
aining process. Our model improves the state-of-the-art performance by 0.5 BLEU
points on the WMT 2014 English-to-German translation task and by 0.4 on the Engl
ish-to-French translation task.
****
Integrating Episodic Memory into a Reinforcement Learning Agent Using Reservoir
Sampling
Kenny J. Young,Shuo Yang,Richard S. Sutton
Episodic memory is a psychology term which refers to the ability to recall speci
fic events from the past. We suggest one advantage of this particular type of me
mory is the ability to easily assign credit to a specific state when remembered
information is found to be useful. Inspired by this idea, and the increasing pop
ularity of external memory mechanisms to handle long-term dependencies in deep l
earning systems, we propose a novel algorithm which uses a reservoir sampling pr

ocedure to maintain an external memory consisting of a fixed number of past stat
es. The algorithm allows a deep reinforcement learning agent to learn online to
preferentially remember those states which are found to be useful to recall late
r on. Critically this method allows for efficient online computation of gradient
 estimates with respect to the write process of the external memory. Thus unlike
 most prior mechanisms for external memory it is feasible to use in an online re
inforcement learning setting.

****
Decoding Decoders: Finding Optimal Representation Spaces for Unsupervised Simila
rity Tasks
Vitalii Zhelezniak,Dan Busbridge,April Shen,Samuel L. Smith,Nils Y. Hammerla
Experimental evidence indicates that simple models outperform complex deep netwo
rks on many unsupervised similarity tasks. Introducing the concept of an optimal
 representation space, we provide a simple theoretical resolution to this appare
nt paradox. In addition, we present a straightforward procedure that, without an
y retraining or architectural modifications, allows deep recurrent models to per
form equally well (and sometimes better) when compared to shallow models. To val
idate our analysis, we conduct a set of consistent empirical evaluations and int
roduce several new sentence embedding models in the process. Even though this wo
rk is presented within the context of natural language processing, the insights
are readily applicable to other domains that rely on distributed representations
 for transfer tasks.
****
Representing Entropy : A short proof of the equivalence between soft Q-learning
and policy gradients
Pierre H. Richemond,Brendan Maginnis
Two main families of reinforcement learning algorithms, Q-learning and policy gr
adients, have recently been proven to be equivalent when using a softmax relaxat
ion on one part, and an entropic regularization on the other. We relate this res
ult to the well-known convex duality of Shannon entropy and the softmax function
. Such a result is also known as the Donsker-Varadhan formula. This provides a s
hort proof of the equivalence. We then interpret this duality further, and use i
deas of convex analysis to prove a new policy inequality relative to soft Q-lear
ning.
****
Skip Connections Eliminate Singularities
Emin Orhan,Xaq Pitkow
Skip connections made the training of very deep networks possible and have becom
e an indispensable component in a variety of neural architectures. A completely
satisfactory explanation for their success remains elusive. Here, we present a n
ovel explanation for the benefits of skip connections in training very deep netw
orks. The difficulty of training deep networks is partly due to the singularitie
s caused by the non-identifiability of the model. Several such singularities hav
e been identified in previous works: (i) overlap singularities caused by the per
mutation symmetry of nodes in a given layer, (ii) elimination singularities corr
esponding to the elimination, i.e. consistent deactivation, of nodes, (iii) sing
ularities generated by the linear dependence of the nodes. These singularities c
ause degenerate manifolds in the loss landscape that slow down learning. We argu
e that skip connections eliminate these singularities by breaking the permutatio
n symmetry of nodes, by reducing the possibility of node elimination and by maki
ng the nodes less linearly dependent. Moreover, for typical initializations, ski
p connections move the network away from the "ghosts" of these singularities and
 sculpt the landscape around them to alleviate the learning slow-down. These hyp
otheses are supported by evidence from simplified models, as well as from experi
ments with deep networks trained on real-world datasets.
****
Deep learning mutation prediction enables early stage lung cancer detection in l
iquid biopsy
Steven T. Kothen-Hill,Asaf Zviran,Rafael C. Schulman,Sunil Deochand,Federico Gai

ti,Dillon Maloney,Kevin Y. Huang,Will Liao,Nicolas Robine,Nathaniel D. Omans,Dan A. Landau
Somatic cancer mutation detection at ultra-low variant allele frequencies (VAFs) is an unmet challenge that is intractable with current state-of-the-art mutation calling methods. Specifically, the limit of VAF detection is closely related to the depth of coverage, due to the requirement of multiple supporting reads in extant methods, precluding the detection of mutations at VAFs that are orders of magnitude lower than the depth of coverage. Nevertheless, the ability to detect cancer-associated mutations in ultra low VAFs is a fundamental requirement for low-tumor burden cancer diagnostics applications such as early detection, monitoring, and therapy nomination using liquid biopsy methods (cell-free DNA). Here we defined a spatial representation of sequencing information adapted for convolutional architecture that enables variant detection at VAFs, in a manner independent of the depth of sequencing. This method enables the detection of cancer mutations even in VAFs as low as 10x-4^, >2 orders of magnitude below the current state-of-the-art. We validated our method on both simulated plasma and on clinical cfDNA plasma samples from cancer patients and non-cancer controls. This method introduces a new domain within bioinformatics and personalized medicine – somatic whole genome mutation calling for liquid biopsy.
****
Thinking like a machine — generating visual rationales through latent space optimization
Jarrel Seah,Jennifer Tang,Andy Kitchen,Jonathan Seah
Interpretability and small labelled datasets are key issues in the practical application of deep learning, particularly in areas such as medicine. In this paper, we present a semi-supervised technique that addresses both these issues simultaneously. We learn dense representations from large unlabelled image datasets, then use those representations to both learn classifiers from small labeled sets and generate visual rationales explaining the predictions. Using chest radiography diagnosis as a motivating application, we show our method has good generalization ability by learning to represent our chest radiography dataset while training a classifier on an separate set from a different institution. Our method identifies heart failure and other thoracic diseases. For each prediction, we generate visual rationales for positive classifications by optimizing a latent representation to minimize the probability of disease while constrained by a similarity measure in image space. Decoding the resultant latent representation produces an image without apparent disease. The difference between the original and the altered image forms an interpretable visual rationale for the algorithm's prediction. Our method simultaneously produces visual rationales that compare favourably to previous techniques and a classifier that outperforms the current state-of-the-art.
****
Distributed Prioritized Experience Replay
Dan Horgan,John Quan,David Budden,Gabriel Barth-Maron,Matteo Hessel,Hado van Hasselt,David Silver
We propose a distributed architecture for deep reinforcement learning at scale, that enables agents to learn effectively from orders of magnitude more data than previously possible. The algorithm decouples acting from learning: the actors interact with their own instances of the environment by selecting actions according to a shared neural network, and accumulate the resulting experience in a shared experience replay memory; the learner replays samples of experience and updates the neural network. The architecture relies on prioritized experience replay to focus only on the most significant data generated by the actors. Our architecture substantially improves the state of the art on the Arcade Learning Environment, achieving better final performance in a fraction of the wall-clock training time.
****
Video Action Segmentation with Hybrid Temporal Networks
Li Ding,Chenliang Xu
Action segmentation as a milestone towards building automatic systems to underst

and untrimmed videos has received considerable attention in the recent years. It is typically being modeled as a sequence labeling problem but contains intrinsic and sufficient differences than text parsing or speech processing. In this paper, we introduce a novel hybrid temporal convolutional and recurrent network (TricorNet), which has an encoder-decoder architecture: the encoder consists of a hierarchy of temporal convolutional kernels that capture the local motion changes of different actions; the decoder is a hierarchy of recurrent neural networks that are able to learn and memorize long-term action dependencies after the encoding stage. Our model is simple but extremely effective in terms of video sequence labeling. The experimental results on three public action segmentation datasets have shown that the proposed model achieves superior performance over the state of the art.
****

Variational Bi-LSTMs

Samira Shabanian,Devansh Arpit,Adam Trischler,Yoshua Bengio

Recurrent neural networks like long short-term memory (LSTM) are important architectures for sequential prediction tasks. LSTMs (and RNNs in general) model sequences along the forward time direction. Bidirectional LSTMs (Bi-LSTMs), which model sequences along both forward and backward directions, generally perform better at such tasks because they capture a richer representation of the data. In the training of Bi-LSTMs, the forward and backward paths are learned independently. We propose a variant of the Bi-LSTM architecture, which we call Variational Bi-LSTM, that creates a dependence between the two paths (during training, but which may be omitted during inference). Our model acts as a regularizer and encourages the two networks to inform each other in making their respective predictions using distinct information. We perform ablation studies to better understand the different components of our model and evaluate the method on various benchmarks, showing state-of-the-art performance.
****

Understanding Deep Learning Generalization by Maximum Entropy

Guanhua Zheng,Jitao Sang,Changsheng Xu

Deep learning achieves remarkable generalization capability with overwhelming number of model parameters. Theoretical understanding of deep learning generalization receives recent attention yet remains not fully explored. This paper attempts to provide an alternative understanding from the perspective of maximum entropy. We first derive two feature conditions that softmax regression strictly apply maximum entropy principle. DNN is then regarded as approximating the feature conditions with multilayer feature learning, and proved to be a recursive solution towards maximum entropy principle. The connection between DNN and maximum entropy well explains why typical designs such as shortcut and regularization improves model generalization, and provides instructions for future model development.
****

BinaryFlex: On-the-Fly Kernel Generation in Binary Convolutional Networks

Vincent W.-S. Tseng,Sourav Bhattachary,Javier Fernández Marqués,Milad Alizadeh,Catherine Tong,Nicholas Donald Lane

In this work we present BinaryFlex, a neural network architecture that learns weighting coefficients of predefined orthogonal binary basis instead of the conventional approach of learning directly the convolutional filters. We have demonstrated the feasibility of our approach for complex computer vision datasets such as ImageNet. Our architecture trained on ImageNet is able to achieve top-5 accuracy of 65.7% while being around 2x smaller than binary networks capable of achieving similar accuracy levels. By using deterministic basis, that can be generated on-the-fly very efficiently, our architecture offers a great deal of flexibility in memory footprint when deploying in constrained microcontroller devices.
****

Associative Conversation Model: Generating Visual Information from Textual Information

Yoichi Ishibashi,Hisashi Miyamori

In this paper, we propose the Associative Conversation Model that generates visual information from textual information and uses it for generating sentences in

order to utilize visual information in a dialogue system without image input. In research on Neural Machine Translation, there are studies that generate translated sentences using both images and sentences, and these studies show that visual information improves translation performance. However, it is not possible to use sentence generation algorithms using images for the dialogue systems since many text-based dialogue systems only accept text input. Our approach generates (associates) visual information from input text and generates response text using context vector fusing associative visual information and sentence textual information. A comparative experiment between our proposed model and a model without association showed that our proposed model is generating useful sentences by associating visual information related to sentences. Furthermore, analysis experiment of visual association showed that our proposed model generates (associates) visual information effective for sentence generation.
****

Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks
Shiyu Liang,Yixuan Li,R. Srikant
We consider the problem of detecting out-of-distribution images in neural networks. We propose ODIN, a simple and effective method that does not require any change to a pre-trained neural network. Our method is based on the observation that using temperature scaling and adding small perturbations to the input can separate the softmax score distributions of in- and out-of-distribution images, allowing for more effective detection. We show in a series of experiments that ODIN is compatible with diverse network architectures and datasets. It consistently outperforms the baseline approach by a large margin, establishing a new state-of-the-art performance on this task. For example, ODIN reduces the false positive rate from the baseline 34.7% to 4.3% on the DenseNet (applied to CIFAR-10 and Tiny-ImageNet) when the true positive rate is 95%.
****

Continuous Convolutional Neural Networks for Image Classification
Vitor Guizilini,Fabio Ramos
This paper introduces the concept of continuous convolution to neural networks and deep learning applications in general. Rather than directly using discretized information, input data is first projected into a high-dimensional Reproducing Kernel Hilbert Space (RKHS), where it can be modeled as a continuous function using a series of kernel bases. We then proceed to derive a closed-form solution to the continuous convolution operation between two arbitrary functions operating in different RKHS. Within this framework, convolutional filters also take the form of continuous functions, and the training procedure involves learning the RKHS to which each of these filters is projected, alongside their weight parameters. This results in much more expressive filters, that do not require spatial discretization and benefit from properties such as adaptive support and non-stationarity. Experiments on image classification are performed, using classical datasets, with results indicating that the proposed continuous convolutional neural network is able to achieve competitive accuracy rates with far fewer parameters and a faster convergence rate.
****

Grouping-By-ID: Guarding Against Adversarial Domain Shifts
Christina Heinze-Deml,Nicolai Meinshausen
When training a deep neural network for supervised image classification, one can broadly distinguish between two types of latent features of images that will drive the classification of class Y. Following the notation of Gong et al. (2016), we can divide features broadly into the classes of (i) “core” or “conditionally invariant” features X^ci whose distribution P(X^ci | Y) does not change substantially across domains and (ii) “style” or “orthogonal” features X^orth whose distribution P(X^orth | Y) can change substantially across domains. These latter orthogonal features would generally include features such as position, rotation, image quality or brightness but also more complex ones like hair color or posture for images of persons. We try to guard against future adversarial domain shifts by ideally just using the “conditionally invariant” features for classification

. In contrast to previous work, we assume that the domain itself is not observed and hence a latent variable. We can hence not directly see the distributional change of features across different domains.

We do assume, however, that we can sometimes observe a so-called identifier or ID variable. We might know, for example, that two images show the same person, with ID referring to the identity of the person. In data augmentation, we generate several images from the same original image, with ID referring to the relevant original image. The method requires only a small fraction of images to have an ID variable.

We provide a causal framework for the problem by adding the ID variable to the model of Gong et al. (2016). However, we are interested in settings where we cannot observe the domain directly and we treat domain as a latent variable. If two or more samples share the same class and identifier, (Y, ID)=(y,i), then we treat those samples as counterfactuals under different style interventions on the orthogonal or style features. Using this grouping-by-ID approach, we regularize the network to provide near constant output across samples that share the same ID by penalizing with an appropriate graph Laplacian. This is shown to substantially improve performance in settings where domains change in terms of image quality, brightness, color changes, and more complex changes such as changes in movement and posture. We show links to questions of interpretability, fairness and transfer learning.
****
MACHINE VS MACHINE: MINIMAX-OPTIMAL DEFENSE AGAINST ADVERSARIAL EXAMPLES
Jihun Hamm
Recently, researchers have discovered that the state-of-the-art object classifiers can be fooled easily by small perturbations in the input unnoticeable to human eyes. It is known that an attacker can generate strong adversarial examples if she knows the classifier parameters. Conversely, a defender can robustify the classifier by retraining if she has the adversarial examples.
The cat-and-mouse game nature of attacks and defenses raises the question of the presence of equilibria in the dynamics.
In this paper, we present a neural-network based attack class to approximate a larger but intractable class of attacks, and
formulate the attacker-defender interaction as a zero-sum leader-follower game.
We present sensitivity-penalized optimization algorithms to find minimax solutions, which are the best worst-case defenses against whitebox attacks. Advantages of the learning-based attacks and defenses compared to gradient-based attacks and defenses are demonstrated with MNIST and CIFAR-10.
****
FastGCN: Fast Learning with Graph Convolutional Networks via Importance Sampling
Jie Chen,Tengfei Ma,Cao Xiao
The graph convolutional networks (GCN) recently proposed by Kipf and Welling are an effective graph model for semi-supervised learning. Such a model, however, is transductive in nature because parameters are learned through convolutions with both training and test data. Moreover, the recursive neighborhood expansion across layers poses time and memory challenges for training with large, dense graphs. To relax the requirement of simultaneous availability of test data, we interpret graph convolutions as integral transforms of embedding functions under probability measures. Such an interpretation allows for the use of Monte Carlo approaches to consistently estimate the integrals, which in turn leads to a batched training scheme as we propose in this work---FastGCN. Enhanced with importance sampling, FastGCN not only is efficient for training but also generalizes well for inference. We show a comprehensive set of experiments to demonstrate its effectiveness compared with GCN and related models. In particular, training is orders of magnitude more efficient while predictions remain comparably accurate.

****
Intriguing Properties of Adversarial Examples

Ekin Dogus Cubuk,Barret Zoph,Samuel Stern Schoenholz,Quoc V. Le
It is becoming increasingly clear that many machine learning classifiers are vulnerable to adversarial examples. In attempting to explain the origin of adversarial examples, previous studies have typically focused on the fact that neural networks operate on high dimensional data, they overfit, or they are too linear. Here we show that distributions of logit differences have a universal functional form. This functional form is independent of architecture, dataset, and training protocol; nor does it change during training. This leads to adversarial error having a universal scaling, as a power-law, with respect to the size of the adversarial perturbation. We show that this universality holds for a broad range of datasets (MNIST, CIFAR10, ImageNet, and random data), models (including state-of-the-art deep networks, linear models, adversarially trained networks, and networks trained on randomly shuffled labels), and attacks (FGSM, step l.l., PGD). Motivated by these results, we study the effects of reducing prediction entropy on adversarial robustness. Finally, we study the effect of network architectures on adversarial sensitivity. To do this, we use neural architecture search with reinforcement learning to find adversarially robust architectures on CIFAR10. Our resulting architecture is more robust to white \emph{and} black box attacks compared to previous attempts.

****

A comparison of second-order methods for deep convolutional neural networks
Patrick H. Chen,Cho-jui Hsieh
Despite many second-order methods have been proposed to train neural networks, most of the results were done on smaller single layer fully connected networks, so we still cannot conclude whether it's useful in training deep convolutional networks. In this study, we conduct extensive experiments to answer the question "whether second-order method is useful for deep learning?". In our analysis, we find out although currently second-order methods are too slow to be applied in practice, it can reduce training loss in fewer number of iterations compared with SGD. In addition, we have the following interesting findings: (1) When using a large batch size, inexact-Newton methods will converge much faster than SGD. Therefore inexact-Newton method could be a better choice in distributed training of deep networks. (2) Quasi-newton methods are competitive with SGD even when using ReLu activation function (which has no curvature) on residual networks. However, current methods are too sensitive to parameters and not easy to tune for different settings. Therefore, quasi-newton methods with more self-adjusting mechanisms might be more useful than SGD in training deeper networks.

****

A closer look at the word analogy problem
Siddharth Krishna Kumar
Although word analogy problems have become a standard tool for evaluating word vectors, little is known about why word vectors are so good at solving these problems. In this paper, I attempt to further our understanding of the subject, by developing a simple, but highly accurate generative approach to solve the word analogy problem for the case when all terms involved in the problem are nouns. My results demonstrate the ambiguities associated with learning the relationship between a word pair, and the role of the training dataset in determining the relationship which gets most highlighted. Furthermore, my results show that the ability of a model to accurately solve the word analogy problem may not be indicative of a model’s ability to learn the relationship between a word pair the way a human does.

****

A Simple Fully Connected Network for Composing Word Embeddings from Characters
Michael Traynor,Thomas Trappenberg
This work introduces a simple network for producing character aware word embeddings. Position agnostic and position aware character embeddings are combined to produce an embedding vector for each word. The learned word representations are s

hown to be very sparse and facilitate improved results on language modeling task s, despite using markedly fewer parameters, and without the need to apply dropou t. A final experiment suggests that weight sharing contributes to sparsity, incr eases performance, and prevents overfitting.
****

Kronecker Recurrent Units
Cijo Jose,Moustapha Cisse,Francois Fleuret
Our work addresses two important issues with recurrent neural networks: (1) they are over-parameterized, and (2) the recurrent weight matrix is ill-conditioned. The former increases the sample complexity of learning and the training time. T he latter causes the vanishing and exploding gradient problem. We present a flex ible recurrent neural network model called Kronecker Recurrent Units (KRU). KRU achieves parameter efficiency in RNNs through a Kronecker factored recurrent mat rix. It overcomes the ill-conditioning of the recurrent matrix by enforcing soft unitary constraints on the factors. Thanks to the small dimensionality of the f actors, maintaining these constraints is computationally efficient. Our experime ntal results on seven standard data-sets reveal that KRU can reduce the number o f parameters by three orders of magnitude in the recurrent weight matrix compare d to the existing recurrent models, without trading the statistical performance. These results in particular show that while there are advantages in having a hi gh dimensional recurrent space, the capacity of the recurrent part of the model can be dramatically reduced.
****

Model-Ensemble Trust-Region Policy Optimization
Thanard Kurutach,Ignasi Clavera,Yan Duan,Aviv Tamar,Pieter Abbeel
Model-free reinforcement learning (RL) methods are succeeding in a growing numbe r of tasks, aided by recent advances in deep learning. However, they tend to su ffer from high sample complexity, which hinders their use in real-world domains. Alternatively, model-based reinforcement learning promises to reduce sample co mplexity, but tends to require careful tuning and to date have succeeded mainly in restrictive domains where simple models are sufficient for learning. In this paper, we analyze the behavior of vanilla model-based reinforcement learning met hods when deep neural networks are used to learn both the model and the policy, and show that the learned policy tends to exploit regions where insufficient dat a is available for the model to be learned, causing instability in training. To overcome this issue, we propose to use an ensemble of models to maintain the mod el uncertainty and regularize the learning process. We further show that the use of likelihood ratio derivatives yields much more stable learning than backpropa gation through time. Altogether, our approach Model-Ensemble Trust-Region Policy Optimization (ME-TRPO) significantly reduces the sample complexity compared to model-free deep RL methods on challenging continuous control benchmark tasks.
****

Deep Sensing: Active Sensing using Multi-directional Recurrent Neural Networks
Jinsung Yoon,William R. Zame,Mihaela van der Schaar
For every prediction we might wish to make, we must decide what to observe (what source of information) and when to observe it. Because making observations is c ostly, this decision must trade off the value of information against the cost of observation. Making observations (sensing) should be an active choice. To solve the problem of active sensing we develop a novel deep learning architecture: De ep Sensing. At training time, Deep Sensing learns how to issue predictions at va rious cost-performance points. To do this, it creates multiple representations a t various performance levels associated with different measurement rates (costs) . This requires learning how to estimate the value of real measurements vs. infe rred measurements, which in turn requires learning how to infer missing (unobser ved) measurements. To infer missing measurements, we develop a Multi-directiona l Recurrent Neural Network (M-RNN). An M-RNN differs from a bi-directional RNN in that it sequentially operates across streams in addition to within streams, a nd because the timing of inputs into the hidden layers is both lagged and advanc ed. At runtime, the operator prescribes a performance level or a cost constrain t, and Deep Sensing determines what measurements to take and what to infer from

those measurements, and then issues predictions. To demonstrate the power of our method, we apply it to two real-world medical datasets with significantly improved performance.
****

PACT: Parameterized Clipping Activation for Quantized Neural Networks
Jungwook Choi,Zhuo Wang,Swagath Venkataramani,Pierce I-Jen Chuang,Vijayalakshmi Srinivasan,Kailash Gopalakrishnan
Deep learning algorithms achieve high classification accuracy at the expense of significant computation cost. To address this cost, a number of quantization schemeshave been proposed - but most of these techniques focused on quantizing weights, which are relatively smaller in size compared to activations. This paper proposes a novel quantization scheme for activations during training - that enables neural networks to work well with ultra low precision weights and activations without any significant accuracy degradation. This technique, PArameterized Clipping acTi-vation (PACT), uses an activation clipping parameter α that is optimized duringtraining to find the right quantization scale. PACT allows quantizing activations toarbitrary bit precisions, while achieving much better accuracy relative to publishedstate-of-the-art quantization schemes. We show, for the first time, that both weights and activations can be quantized to 4-bits of precision while still achieving accuracy comparable to full precision networks across a range of popular models and datasets. We also show that exploiting these reduced-precision computational units in hardware can enable a super-linear improvement in inferencing performance dueto a significant reduction in the area of accelerator compute engines coupled with the ability to retain the quantized model and activation data in on-chip memories.
****

Lifelong Word Embedding via Meta-Learning
Hu Xu,Bing Liu,Lei Shu,Philip S. Yu
Learning high-quality word embeddings is of significant importance in achieving better performance in many down-stream learning tasks. On one hand, traditional word embeddings are trained on a large scale corpus for general-purpose tasks, which are often sub-optimal for many domain-specific tasks. On the other hand, many domain-specific tasks do not have a large enough domain corpus to obtain high-quality embeddings. We observe that domains are not isolated and a small domain corpus can leverage the learned knowledge from many past domains to augment that corpus in order to generate high-quality embeddings. In this paper, we formulate the learning of word embeddings as a lifelong learning process. Given knowledge learned from many previous domains and a small new domain corpus, the proposed method can effectively generate new domain embeddings by leveraging a simple but effective algorithm and a meta-learner, where the meta-learner is able to provide word context similarity information at the domain-level. Experimental results demonstrate that the proposed method can effectively learn new domain embeddings from a small corpus and past domain knowledges\footnote{We will release the code after final revisions.}. We also demonstrate that general-purpose embeddings trained from a large scale corpus are sub-optimal in domain-specific tasks.
****

Graph Attention Networks
Petar Veli■kovi■,Guillem Cucurull,Arantxa Casanova,Adriana Romero,Pietro Liò,Yoshua Bengio
We present graph attention networks (GATs), novel neural network architectures that operate on graph-structured data, leveraging masked self-attentional layers to address the shortcomings of prior methods based on graph convolutions or their approximations. By stacking layers in which nodes are able to attend over their neighborhoods' features, we enable (implicitly) specifying different weights to different nodes in a neighborhood, without requiring any kind of computationally intensive matrix operation (such as inversion) or depending on knowing the graph structure upfront. In this way, we address several key challenges of spectral-based graph neural networks simultaneously, and make our model readily applicable to inductive as well as transductive problems. Our GAT models have achieved or matched state-of-the-art results across four established transductive and ind

uctive graph benchmarks: the Cora, Citeseer and Pubmed citation network datasets, as well as a protein-protein interaction dataset (wherein test graphs remain unseen during training).
****

UNSUPERVISED SENTENCE EMBEDDING USING DOCUMENT STRUCTURE-BASED CONTEXT
Taesung Lee,Youngja Park
We present a new unsupervised method for learning general-purpose sentence embeddings.
Unlike existing methods which rely on local contexts, such as words
inside the sentence or immediately neighboring sentences, our method selects, for
each target sentence, influential sentences in the entire document based on a document
structure. We identify a dependency structure of sentences using metadata
or text styles. Furthermore, we propose a novel out-of-vocabulary word handling
technique to model many domain-specific terms, which were mostly discarded by
existing sentence embedding methods. We validate our model on several tasks
showing 30% precision improvement in coreference resolution in a technical domain,
and 7.5% accuracy increase in paraphrase detection compared to baselines.
****

CyCADA: Cycle-Consistent Adversarial Domain Adaptation
Judy Hoffman,Eric Tzeng,Taesung Park,Jun-Yan Zhu,Phillip Isola,Kate Saenko,Alyosha Efros,Trevor Darrell
Domain adaptation is critical for success in new, unseen environments.
Adversarial adaptation models applied in feature spaces discover domain invariant representations, but are difficult to visualize and sometimes fail to capture pixel-level and low-level domain shifts.
Recent work has shown that generative adversarial networks combined with cycle-consistency constraints are surprisingly effective at mapping images between domains, even without the use of aligned image pairs.
We propose a novel discriminatively-trained Cycle-Consistent Adversarial Domain Adaptation model.
CyCADA adapts representations at both the pixel-level and feature-level, enforces cycle-consistency while leveraging a task loss, and does not require aligned pairs. Our model can be applied in a variety of visual recognition and prediction settings.
We show new state-of-the-art results across multiple adaptation tasks, including digit classification and semantic segmentation of road scenes demonstrating transfer from synthetic to real world domains.
****

Don't encrypt the data; just approximate the model \ Towards Secure Transaction and Fair Pricing of Training Data
Xinlei Xu
As machine learning becomes ubiquitous, deployed systems need to be as accu- rate as they can. As a result, machine learning service providers have a surging need for useful, additional training data that benefits training, without giving up all the details about the trained program. At the same time, data owners would like to trade their data for its value, without having to first give away the data itself be- fore receiving compensation. It is difficult for data providers and model providers to agree on a fair price without first revealing the data or the trained model to the other side. Escrow systems only complicate this further, adding an additional layer of trust required of both parties. Currently, data owners and model owners don’t have a fair pricing system that eliminates the need to trust a third party and training the model on the data, which 1) takes a long time to complete, 2) does not guarantee that useful data is paid valuably and that useless data isn’t, without trusting in the third party with both the model and the data. Existing improve- ments to secure the transaction focus heavily on encrypting or approximating the data, such as training on encrypted data, and variants of federated learning. As powerful as the methods appear to be, we sho

w them to be impractical in our use case with real world assumptions for preserv ing privacy for the data owners when facing black-box models. Thus, a fair prici ng scheme that does not rely on secure data encryption and obfuscation is needed before the exchange of data. This pa- per proposes a novel method for fair pric ing using data-model efficacy techniques such as influence functions, model extr action, and model compression methods, thus enabling secure data transactions. W e successfully show that without running the data through the model, one can app roximate the value of the data; that is, if the data turns out redundant, the pr icing is minimal, and if the data leads to proper improvement, its value is prop erly assessed, without placing strong assumptions on the nature of the model. Fu ture work will be focused on establishing a system with stronger transactional s ecurity against adversarial attacks that will reveal details about the model or the data to the other party.
****
Network Iterative Learning for Dynamic Deep Neural Networks via Morphism
Tao Wei,Changhu Wang,Chang Wen Chen
In this research, we present a novel learning scheme called network iterative le arning for deep neural networks. Different from traditional optimization algorit hms that usually optimize directly on a static objective function, we propose in this work to optimize a dynamic objective function in an iterative fashion capa ble of adapting its function form when being optimized. The optimization is impl emented as a series of intermediate neural net functions that is able to dynamic ally grow into the targeted neural net objective function. This is done via netw ork morphism so that the network knowledge is fully preserved with each network growth. Experimental results demonstrate that the proposed network iterative lea rning scheme is able to significantly alleviate the degradation problem. Its eff ectiveness is verified on diverse benchmark datasets.
****
Learning Graph Convolution Filters from Data Manifold
Guokun Lai,Hanxiao Liu,Yiming Yang
Convolution Neural Network (CNN) has gained tremendous success in computer visio n tasks with its outstanding ability to capture the local latent features. Recen tly, there has been an increasing interest in extending CNNs to the general spat ial domain. Although various types of graph convolution and geometric convolutio n methods have been proposed, their connections to traditional 2D-convolution ar e not well-understood. In this paper, we show that depthwise separable convoluti on is a path to unify the two kinds of convolution methods in one mathematical v iew, based on which we derive a novel Depthwise Separable Graph Convolution that subsumes existing graph convolution methods as special cases of our formulation . Experiments show that the proposed approach consistently outperforms other gra ph convolution and geometric convolution baselines on benchmark datasets in mult iple domains.
****
Compact Neural Networks based on the Multiscale Entanglement Renormalization Ans atz
Andrew Hallam,Edward Grant,Vid Stojevic,Simone Severini,Andrew G. Green
The goal of this paper is to demonstrate a method for tensorizing neural network s based upon an efficient way of approximating scale invariant quantum states, t he Multi-scale Entanglement Renormalization Ansatz (MERA). We employ MERA as a r eplacement for linear layers in a neural network and test this implementation on the CIFAR-10 dataset. The proposed method outperforms factorization using tenso r trains, providing greater compression for the same level of accuracy and great er accuracy for the same level of compression. We demonstrate MERA-layers with 3 900 times fewer parameters and a reduction in accuracy of less than 1% compared to the equivalent fully connected layers.

****
A Self-Training Method for Semi-Supervised GANs
Alan Do-Omri,Dalei Wu,Xiaohua Liu
Since the creation of Generative Adversarial Networks (GANs), much work has been

 done to improve their training stability, their generated image quality, their
range of application but nearly none of them explored their self-training potent
ial. Self-training has been used before the advent of deep learning in order to
allow training on limited labelled training data and has shown impressive result
s in semi-supervised learning. In this work, we combine these two ideas and make
 GANs self-trainable for semi-supervised learning tasks by exploiting their infi
nite data generation potential. Results show that using even the simplest form o
f self-training yields an improvement. We also show results for a more complex s
elf-training scheme that performs at least as well as the basic self-training sc
heme but with significantly less data augmentation.
****
XGAN: Unsupervised Image-to-Image Translation for many-to-many Mappings
Amelie Royer,Konstantinos Bousmalis,Stephan Gouws,Fred Bertsch,Inbar Mosseri,For
rester Cole,Kevin Murphy
Style transfer usually refers to the task of applying color and texture informat
ion from a specific style image to a given content image while preserving the st
ructure of the latter. Here we tackle the more generic problem of semantic style
 transfer: given two unpaired collections of images, we aim to learn a mapping b
etween the corpus-level style of each collection, while preserving semantic cont
ent shared across the two domains. We introduce XGAN ("Cross-GAN"), a dual adver
sarial autoencoder, which captures a shared representation of the common domain
semantic content in an unsupervised way, while jointly learning the domain-to-do
main image translations in both directions. We exploit ideas from the domain ad
aptation literature and define a semantic consistency loss which encourages the
model to preserve semantics in the learned embedding space. We report promising
qualitative results for the task of face-to-cartoon translation. The cartoon dat
aset we collected for this purpose will also be released as a new benchmark for
semantic style transfer.
****
Unsupervised Representation Learning by Predicting Image Rotations
Spyros Gidaris,Praveer Singh,Nikos Komodakis
Over the last years, deep convolutional neural networks (ConvNets) have transfor
med the field of computer vision thanks to their unparalleled capacity to learn
 high level semantic image features. However, in order to successfully learn tho
se features, they usually require massive amounts of manually labeled data, whic
h is both expensive and impractical to scale. Therefore, unsupervised semantic f
eature learning, i.e., learning without requiring manual annotation effort, is o
f crucial importance in order to successfully harvest the vast amount of visual
data that are available today. In our work we propose to learn image features by
 training ConvNets to recognize the 2d rotation that is applied to the image tha
t it gets as input. We demonstrate both qualitatively and quantitatively that t
his apparently simple task actually provides a very powerful supervisory signal
for semantic feature learning. We exhaustively evaluate our method in various u
nsupervised feature learning benchmarks and we exhibit in all of them state-of-t
he-art performance. Specifically, our results on those benchmarks demonstrate dr
amatic improvements w.r.t. prior state-of-the-art approaches in unsupervised rep
resentation learning and thus significantly close the gap with supervised featur
e learning. For instance, in PASCAL VOC 2007 detection task our unsupervised pre
-trained AlexNet model achieves the state-of-the-art (among unsupervised methods
) mAP of 54.4%$that is only 2.4 points lower from the supervised case. We get s
imilarly striking results when we transfer our unsupervised learned features on
various other tasks, such as ImageNet classification, PASCAL classification, PAS
CAL segmentation, and CIFAR-10 classification. The code and models of our paper
will be published on:
https://github.com/gidariss/FeatureLearningRotNet
****
The Role of Minimal Complexity Functions in Unsupervised Learning of Semantic Ma
ppings
Tomer Galanti,Lior Wolf,Sagie Benaim
We discuss the feasibility of the following learning problem: given unmatched sa

mples from two domains and nothing else, learn a mapping between the two, which
preserves semantics. Due to the lack of paired samples and without any definitio
n of the semantic information, the problem might seem ill-posed. Specifically, i
n typical cases, it seems possible to build infinitely many alternative mappings
 from every target mapping. This apparent ambiguity stands in sharp contrast to
 the recent empirical success in solving this problem.

We identify the abstract notion of aligning two domains in a semantic way with c
oncrete terms of minimal relative complexity. A theoretical framework for measur
ing the complexity of compositions of functions is developed in order to show th
at it is reasonable to expect the minimal complexity mapping to be unique. The m
easured complexity used is directly related to the depth of the neural networks
being learned and a semantically aligned mapping could then be captured simply b
y learning using architectures that are not much bigger than the minimal archite
cture.

Various predictions are made based on the hypothesis that semantic alignment can
 be captured by the minimal mapping. These are verified extensively. In addition
, a new mapping algorithm is proposed and shown to lead to better mapping result
s.
****
The Cramer Distance as a Solution to Biased Wasserstein Gradients
Marc G. Bellemare,Ivo Danihelka,Will Dabney,Shakir Mohamed,Balaji Lakshminarayan
an,Stephan Hoyer,Remi Munos
The Wasserstein probability metric has received much attention from the machine
learning community. Unlike the Kullback-Leibler divergence, which strictly measu
res change in probability, the Wasserstein metric reflects the underlying geomet
ry between outcomes. The value of being sensitive to this geometry has been demo
nstrated, among others, in ordinal regression and generative modelling, and most
 recently in reinforcement learning. In this paper we describe three natural pro
perties of probability divergences that we believe reflect requirements from mac
hine learning: sum invariance, scale sensitivity, and unbiased sample gradients.
 The Wasserstein metric possesses the first two properties but, unlike the Kullb
ack-Leibler divergence, does not possess the third. We provide empirical evidenc
e suggesting this is a serious issue in practice. Leveraging insights from proba
bilistic forecasting we propose an alternative to the Wasserstein metric, the Cr
amér distance. We show that the Cramér distance possesses all three desired prop
erties, combining the best of the Wasserstein and Kullback-Leibler divergences.
We give empirical results on a number of domains comparing these three divergenc
es. To illustrate the practical relevance of the Cramér distance we design a new
 algorithm, the Cramér Generative Adversarial Network (GAN), and show that it ha
s a number of desirable properties over the related Wasserstein GAN.

****
Rethinking generalization requires revisiting old ideas: statistical mechanics a
pproaches and complex learning behavior
Charles H. Martin,Michael W. Mahoney
We describe an approach to understand the peculiar and counterintuitive generali
zation properties of deep neural networks. The approach involves going beyond w
orst-case theoretical capacity control frameworks that have been popular in mach
ine learning in recent years to revisit old ideas in the statistical mechanics o
f neural networks. Within this approach, we present a prototypical Very Simple
Deep Learning (VSDL) model, whose behavior is controlled by two control paramete
rs, one describing an effective amount of data, or load, on the network (that de
creases when noise is added to the input), and one with an effective temperature
 interpretation (that increases when algorithms are early stopped). Using this
model, we describe how a very simple application of ideas from the statistical m
echanics theory of generalization provides a strong qualitative description of r
ecently-observed empirical results regarding the inability of deep neural networ
ks not to overfit training data, discontinuous learning and sharp transitions in

the generalization properties of learning algorithms, etc.
****
Evolutionary Expectation Maximization for Generative Models with Binary Latents
Enrico Guiraud,Jakob Drefs,Joerg Luecke
We establish a theoretical link between evolutionary algorithms and variational parameter optimization of probabilistic generative models with binary hidden variables.
While the novel approach is independent of the actual generative model, here we use two such models to investigate its applicability and scalability: a noisy-OR Bayes Net (as a standard example of binary data) and Binary Sparse Coding (as a model for continuous data).

Learning of probabilistic generative models is first formulated as approximate maximum likelihood optimization using variational expectation maximization (EM). We choose truncated posteriors as variational distributions in which discrete latent states serve as variational parameters. In the variational E-step,
the latent states are then
optimized according to a tractable free-energy objective. Given a data point, we can show that evolutionary algorithms can be used for the variational optimization loop by (A)~considering the bit-vectors of the latent states as genomes of individuals, and by (B)~defining the fitness of the
individuals as the (log) joint probabilities given by the used generative model.

As a proof of concept, we apply the novel evolutionary EM approach to the optimization of the parameters of noisy-OR Bayes nets and binary sparse coding on artificial and real data (natural image patches). Using point mutations and single-point cross-over for the evolutionary algorithm, we find that scalable variational EM algorithms are obtained which efficiently improve the data likelihood. In general we believe that, with the link established here, standard as well as recent results in the field of evolutionary optimization can be leveraged to address the difficult problem of parameter optimization in generative models.
****
Learning objects from pixels
David Saxton
We show how discrete objects can be learnt in an unsupervised fashion from pixels, and how to perform reinforcement learning using this object representation.

More precisely, we construct a differentiable mapping from an image to a discrete tabular list of objects, where each object consists of a differentiable position, feature vector, and scalar presence value that allows the representation to be learnt using an attention mechanism.

Applying this mapping to Atari games, together with an interaction net-style architecture for calculating quantities from objects, we construct agents that can play Atari games using objects learnt in an unsupervised fashion. During training, many natural objects emerge, such as the ball and paddles in Pong, and the submarine and fish in Seaquest.

This gives the first reinforcement learning agent for Atari with an interpretable object representation, and opens the avenue for agents that can conduct object-based exploration and generalization.
****
Lifelong Learning with Dynamically Expandable Networks
Jaehong Yoon,Eunho Yang,Jeongtae Lee,Sung Ju Hwang
We propose a novel deep network architecture for lifelong learning which we refer to as Dynamically Expandable Network (DEN), that can dynamically decide its network capacity as it trains on a sequence of tasks, to learn a compact overlapping knowledge sharing structure among tasks. DEN is efficiently trained in an online manner by performing selective retraining, dynamically expands network capacity upon arrival of each task with only the necessary number of units, and effec

tively prevents semantic drift by splitting/duplicating units and timestamping them. We validate DEN on multiple public datasets in lifelong learning scenarios on multiple public datasets, on which it not only significantly outperforms existing lifelong learning methods for deep networks, but also achieves the same level of performance as the batch model with substantially fewer number of parameters.
****
Parallelizing Linear Recurrent Neural Nets Over Sequence Length
Eric Martin,Chris Cundy
Recurrent neural networks (RNNs) are widely used to model sequential data but their non-linear dependencies between sequence elements prevent parallelizing training over sequence length. We show the training of RNNs with only linear sequential dependencies can be parallelized over the sequence length using the parallel scan algorithm, leading to rapid training on long sequences even with small minibatch size. We develop a parallel linear recurrence CUDA kernel and show that it can be applied to immediately speed up training and inference of several state of the art RNN architectures by up to 9x. We abstract recent work on linear RNNs into a new framework of linear surrogate RNNs and develop a linear surrogate model for the long short-term memory unit, the GILR-LSTM, that utilizes parallel linear recurrence. We extend sequence learning to new extremely long sequence regimes that were previously out of reach by successfully training a GILR-LSTM on a synthetic sequence classification task with a one million timestep dependency.

****
Learning Robust Rewards with Adverserial Inverse Reinforcement Learning
Justin Fu,Katie Luo,Sergey Levine
Reinforcement learning provides a powerful and general framework for decision making and control, but its application in practice is often hindered by the need for extensive feature and reward engineering. Deep reinforcement learning methods can remove the need for explicit engineering of policy or value features, but still require a manually specified reward function. Inverse reinforcement learning holds the promise of automatic reward acquisition, but has proven exceptionally difficult to apply to large, high-dimensional problems with unknown dynamics. In this work, we propose AIRL, a practical and scalable inverse reinforcement learning algorithm based on an adversarial reward learning formulation that is competitive with direct imitation learning algorithms. Additionally, we show that AIRL is able to recover portable reward functions that are robust to changes in dynamics, enabling us to learn policies even under significant variation in the environment seen during training.
****
Variance Reduction for Policy Gradient with Action-Dependent Factorized Baselines
Cathy Wu,Aravind Rajeswaran,Yan Duan,Vikash Kumar,Alexandre M Bayen,Sham Kakade,Igor Mordatch,Pieter Abbeel
Policy gradient methods have enjoyed great success in deep reinforcement learning but suffer from high variance of gradient estimates. The high variance problem is particularly exasperated in problems with long horizons or high-dimensional action spaces. To mitigate this issue, we derive a bias-free action-dependent baseline for variance reduction which fully exploits the structural form of the stochastic policy itself and does not make any additional assumptions about the MDP. We demonstrate and quantify the benefit of the action-dependent baseline through both theoretical analysis as well as numerical results, including an analysi

s of the suboptimality of the optimal state-dependent baseline. The result is a computationally efficient policy gradient algorithm, which scales to high-dimensional control problems, as demonstrated by a synthetic 2000-dimensional target matching task. Our experimental results indicate that action-dependent baselines allow for faster learning on standard reinforcement learning benchmarks and high-dimensional hand manipulation and synthetic tasks. Finally, we show that the general idea of including additional information in baselines for improved variance reduction can be extended to partially observed and multi-agent tasks.
****

Reinforcement Learning Algorithm Selection

Romain Laroche,Raphael Feraud

This paper formalises the problem of online algorithm selection in the context of Reinforcement Learning (RL). The setup is as follows: given an episodic task and a finite number of off-policy RL algorithms, a meta-algorithm has to decide which RL algorithm is in control during the next episode so as to maximize the expected return. The article presents a novel meta-algorithm, called Epochal Stochastic Bandit Algorithm Selection (ESBAS). Its principle is to freeze the policy updates at each epoch, and to leave a rebooted stochastic bandit in charge of the algorithm selection. Under some assumptions, a thorough theoretical analysis demonstrates its near-optimality considering the structural sampling budget limitations. ESBAS is first empirically evaluated on a dialogue task where it is shown to outperform each individual algorithm in most configurations. ESBAS is then adapted to a true online setting where algorithms update their policies after each transition, which we call SSBAS. SSBAS is evaluated on a fruit collection task where it is shown to adapt the stepsize parameter more efficiently than the classical hyperbolic decay, and on an Atari game, where it improves the performance by a wide margin.
****

Learning Efficient Tensor Representations with Ring Structure Networks

Qibin Zhao,Masashi Sugiyama,Longhao Yuan,Andrzej Cichocki

\emph{Tensor train (TT) decomposition} is a powerful representation for high-order tensors, which has been successfully applied to various machine learning tasks in recent years. In this paper, we propose a more generalized tensor decomposition with ring structure network by employing circular multilinear products over a sequence of lower-order core tensors, which is termed as TR representation. Several learning algorithms including blockwise ALS with adaptive tensor ranks and SGD with high scalability are presented. Furthermore, the mathematical properties are investigated, which enables us to perform basic algebra operations in a computationally efficiently way by using TR representations. Experimental results on synthetic signals and real-world datasets demonstrate the effectiveness of TR model and the learning algorithms. In particular, we show that the structure information and high-order correlations within a 2D image can be captured efficiently by employing tensorization and TR representation.

****

Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models

Wieland Brendel *,Jonas Rauber *,Matthias Bethge

Many machine learning algorithms are vulnerable to almost imperceptible perturbations of their inputs. So far it was unclear how much risk adversarial perturbations carry for the safety of real-world machine learning applications because most methods used to generate such perturbations rely either on detailed model information (gradient-based attacks) or on confidence scores such as class probabilities (score-based attacks), neither of which are available in most real-world scenarios. In many such cases one currently needs to retreat to transfer-based attacks which rely on cumbersome substitute models, need access to the training data and can be defended against. Here we emphasise the importance of attacks which solely rely on the final model decision. Such decision-based attacks are (1) applicable to real-world black-box models such as autonomous cars, (2) need less knowledge and are easier to apply than transfer-based attacks and (3) are more r

obust to simple defences than gradient- or score-based attacks. Previous attacks in this category were limited to simple models or simple datasets. Here we introduce the Boundary Attack, a decision-based attack that starts from a large adversarial perturbation and then seeks to reduce the perturbation while staying adversarial. The attack is conceptually simple, requires close to no hyperparameter tuning, does not rely on substitute models and is competitive with the best gradient-based attacks in standard computer vision tasks like ImageNet. We apply the attack on two black-box algorithms from Clarifai.com. The Boundary Attack in particular and the class of decision-based attacks in general open new avenues to study the robustness of machine learning models and raise new questions regarding the safety of deployed machine learning systems. An implementation of the attack is available as part of Foolbox (https://github.com/bethgelab/foolbox).
****
“Style” Transfer for Musical Audio Using Multiple Time-Frequency Representations
Shaun Barry,Youngmoo Kim
Neural Style Transfer has become a popular technique for
generating images of distinct artistic styles using convolutional neural networks. This
recent success in image style transfer has raised the question of
whether similar methods can be leveraged to alter the “style” of musical
audio. In this work, we attempt long time-scale high-quality audio transfer
and texture synthesis in the time-domain that captures harmonic,
rhythmic, and timbral elements related to musical style, using examples that
may have different lengths and musical keys. We demonstrate the ability
to use randomly initialized convolutional neural networks to transfer
these aspects of musical style from one piece onto another using 3
different representations of audio: the log-magnitude of the Short Time
Fourier Transform (STFT), the Mel spectrogram, and the Constant-Q Transform
spectrogram. We propose using these representations as a way of
generating and modifying perceptually significant characteristics of
musical audio content. We demonstrate each representation's
shortcomings and advantages over others by carefully designing
neural network structures that complement the nature of musical audio. Finally, we show that the most
compelling “style” transfer examples make use of an ensemble of these
representations to help capture the varying desired characteristics of
audio signals.
****
Nearest Neighbour Radial Basis Function Solvers for Deep Neural Networks
Benjamin J. Meyer,Ben Harwood,Tom Drummond
We present a radial basis function solver for convolutional neural networks that can be directly applied to both distance metric learning and classification problems. Our method treats all training features from a deep neural network as radial basis function centres and computes loss by summing the influence of a feature's nearby centres in the embedding space. Having a radial basis function centred on each training feature is made scalable by treating it as an approximate nearest neighbour search problem. End-to-end learning of the network and solver is carried out, mapping high dimensional features into clusters of the same class. This results in a well formed embedding space, where semantically related instances are likely to be located near one another, regardless of whether or not the network was trained on those classes. The same loss function is used for both the metric learning and classification problems. We show that our radial basis function solver outperforms state-of-the-art embedding approaches on the Stanford Cars196 and CUB-200-2011 datasets. Additionally, we show that when used as a classifier, our method outperforms a conventional softmax classifier on the CUB-200-2011, Stanford Cars196, Oxford 102 Flowers and Leafsnap fine-grained classification datasets.
****
Convolutional Sequence Modeling Revisited
Shaojie Bai,J. Zico Kolter,Vladlen Koltun

This paper revisits the problem of sequence modeling using convolutional architectures. Although both convolutional and recurrent architectures have a long history in sequence prediction, the current "default" mindset in much of the deep learning community is that generic sequence modeling is best handled using recurrent networks. The goal of this paper is to question this assumption.
Specifically, we consider a simple generic temporal convolution network (TCN), which adopts features from modern ConvNet architectures such as a dilations and residual connections. We show that on a variety of sequence modeling tasks, including many frequently used as benchmarks for evaluating recurrent networks, the TCN outperforms baseline RNN methods (LSTMs, GRUs, and vanilla RNNs) and sometimes even highly specialized approaches. We further show that the potential "infinite memory" advantage that RNNs have over TCNs is largely absent in practice: TCNs indeed exhibit longer effective history sizes than their recurrent counterparts. As a whole, we argue that it may be time to (re)consider ConvNets as the default "go to" architecture for sequence modeling.
****
Open Loop Hyperparameter Optimization and Determinantal Point Processes
Jesse Dodge,Kevin Jamieson,Noah A. Smith
Driven by the need for parallelizable hyperparameter optimization methods, this paper studies \emph{open loop} search methods: sequences that are predetermined and can be generated before a single configuration is evaluated. Examples include grid search, uniform random search, low discrepancy sequences, and other sampling distributions.
In particular, we propose the use of k-determinantal point processes in hyperparameter optimization via random search. Compared to conventional uniform random search where hyperparameter settings are sampled independently, a k-DPP promotes diversity. We describe an approach that transforms hyperparameter search spaces for efficient use with a k-DPP. In addition, we introduce a novel Metropolis-Hastings algorithm which can sample from k-DPPs defined over spaces with a mixture of discrete and continuous dimensions. Our experiments show significant benefits over uniform random search in realistic scenarios with a limited budget for training supervised learners, whether in serial or parallel.
****
Generating Wikipedia by Summarizing Long Sequences
Peter J. Liu*,Mohammad Saleh*,Etienne Pot,Ben Goodrich,Ryan Sepassi,Lukasz Kaiser,Noam Shazeer
We show that generating English Wikipedia articles can be approached as a multi-document summarization of source documents. We use extractive summarization to coarsely identify salient information and a neural abstractive model to generate the article. For the abstractive model, we introduce a decoder-only architecture that can scalably attend to very long sequences, much longer than typical encoder-decoder architectures used in sequence transduction. We show that this model can generate fluent, coherent multi-sentence paragraphs and even whole Wikipedia articles. When given reference documents, we show it can extract relevant factual information as reflected in perplexity, ROUGE scores and human evaluations.
****
Variational Message Passing with Structured Inference Networks
Wu Lin,Nicolas Hubacher,Mohammad Emtiyaz Khan
Recent efforts on combining deep models with probabilistic graphical models are promising in providing flexible models that are also easy to interpret. We propose a variational message-passing algorithm for variational inference in such models. We make three contributions. First, we propose structured inference networks that incorporate the structure of the graphical model in the inference network of variational auto-encoders (VAE). Second, we establish conditions under which

 such inference networks enable fast amortized inference similar to VAE. Finally, we derive a variational message passing algorithm to perform efficient natural-gradient inference while retaining the efficiency of the amortized inference. By simultaneously enabling structured, amortized, and natural-gradient inference for deep structured models, our method simplifies and generalizes existing methods.
****
Sobolev GAN
Youssef Mroueh,Chun-Liang Li,Tom Sercu,Anant Raj,Yu Cheng
We propose a new Integral Probability Metric (IPM) between distributions: the Sobolev IPM. The Sobolev IPM compares the mean discrepancy of two distributions for functions (critic) restricted to a Sobolev ball defined with respect to a dominant measure mu. We show that the Sobolev IPM compares two distributions in high dimensions based on weighted conditional Cumulative Distribution Functions (CDF) of each coordinate on a leave one out basis. The Dominant measure mu plays a crucial role as it defines the support on which conditional CDFs are compared. Sobolev IPM can be seen as an extension of the one dimensional Von-Mises Cramer statistics to high dimensional distributions. We show how Sobolev IPM can be used to train Generative Adversarial Networks (GANs). We then exploit the intrinsic conditioning implied by Sobolev IPM in text generation. Finally we show that a variant of Sobolev GAN achieves competitive results in semi-supervised learning on CIFAR-10, thanks to the smoothness enforced on the critic by Sobolev GAN which relates to Laplacian regularization.
****
Adversarial Policy Gradient for Alternating Markov Games
Chao Gao,Martin Mueller,Ryan Hayward
Policy gradient reinforcement learning has been applied to two-player alternate-turn zero-sum games, e.g., in AlphaGo, self-play REINFORCE was used to improve the neural net model after supervised learning. In this paper, we emphasize that two-player zero-sum games with alternating turns, which have been previously formulated as Alternating Markov Games (AMGs), are different from standard MDP because of their two-agent nature. We exploit the difference in associated Bellman equations, which leads to different policy iteration algorithms. As policy gradient method is a kind of generalized policy iteration, we show how these differences in policy iteration are reflected in policy gradient for AMGs. We formulate an adversarial policy gradient and discuss potential possibilities for developing better policy gradient methods other than self-play REINFORCE. The core idea is to estimate the minimum rather than the mean for the “critic”. Experimental results on the game of Hex show the modified Monte Carlo policy gradient methods are able to learn better pure neural net policies than the REINFORCE variants. To apply learned neural weights to multiple board sizes Hex, we describe a board-size independent neural net architecture. We show that when combined with search, using a single neural net model, the resulting program consistently beats MoHex 2.0, the state-of-the-art computer Hex player, on board sizes from 9×9 to 13×13.

****
Hierarchical and Interpretable Skill Acquisition in Multi-task Reinforcement Learning
Tianmin Shu,Caiming Xiong,Richard Socher
Learning policies for complex tasks that require multiple different skills is a major challenge in reinforcement learning (RL). It is also a requirement for its deployment in real-world scenarios. This paper proposes a novel framework for efficient multi-task reinforcement learning. Our framework trains agents to employ hierarchical policies that decide when to use a previously learned policy and when to learn a new skill. This enables agents to continually acquire new skills during different stages of training. Each learned task corresponds to a human language description. Because agents can only access previously learned skills through these descriptions, the agent can always provide a human-interpretable description of its choices. In order to help the agent learn the complex temporal dependencies necessary for the hierarchical policy, we provide it with a stochast

ic temporal grammar that modulates when to rely on previously learned skills and
 when to execute new skills. We validate our approach on Minecraft games designe
d to explicitly test the ability to reuse previously learned skills while simult
aneously learning new skills.
****
Prototype Matching Networks for Large-Scale Multi-label Genomic Sequence Classi
fication
Jack Lanchantin,Arshdeep Sekhon,Ritambhara Singh,Yanjun Qi
One of the fundamental tasks in understanding genomics is the problem of predict
ing Transcription Factor Binding Sites (TFBSs). With more than hundreds of Trans
cription Factors (TFs) as labels, genomic-sequence based TFBS prediction is a ch
allenging multi-label classification task. There are two major biological mechan
isms for TF binding: (1) sequence-specific binding patterns on genomes known as
“motifs” and (2) interactions among TFs known as co-binding effects. In this pap
er, we propose a novel deep architecture, the Prototype Matching Network (PMN) t
o mimic the TF binding mechanisms. Our PMN model automatically extracts prototyp
es (“motif”-like features) for each TF through a novel prototype-matching loss.
Borrowing ideas from few-shot matching models, we use the notion of support set
of prototypes and an LSTM to learn how TFs interact and bind to genomic sequence
s. On a reference TFBS dataset with 2.1 million genomic sequences, PMN significa
ntly outperforms baselines and validates our design choices empirically. To our
knowledge, this is the first deep learning architecture that introduces prototyp
e learning and considers TF-TF interactions for large scale TFBS prediction. Not
 only is the proposed architecture accurate, but it also models the underlying b
iology.
****
When and where do feed-forward neural networks learn localist representations?
Ella M. Gale,Nicolas Martin,Jeffrey Bowers
According to parallel distributed processing (PDP) theory in psychology, neural
networks (NN) learn distributed rather than interpretable localist representatio
ns. This view has been held so strongly that few researchers have analysed singl
e units to determine if this assumption is correct. However, recent results from
 psychology, neuroscience and computer science have shown the occasional existen
ce of local codes emerging in artificial and biological neural networks. In this
 paper, we undertake the first systematic survey of when local codes emerge in a
 feed-forward neural network, using generated input and output data with known q
ualities. We find that the number of local codes that emerge from a NN follows a
 well-defined distribution across the number of hidden layer neurons, with a pea
k determined by the size of input data, number of examples presented and the spa
rsity of input data. Using a 1-hot output code drastically decreases the number
of local codes on the hidden layer. The number of emergent local codes increases
 with the percentage of dropout applied to the hidden layer, suggesting that the
 localist encoding may offer a resilience to noisy networks. This data suggests
that localist coding can emerge from feed-forward PDP networks and suggests some
 of the conditions that may lead to interpretable localist representations in th
e cortex. The findings highlight how local codes should not be dismissed out of
hand.
****
Memory Architectures in Recurrent Neural Network Language Models
Dani Yogatama,Yishu Miao,Gabor Melis,Wang Ling,Adhiguna Kuncoro,Chris Dyer,Phil
Blunsom
We compare and analyze sequential, random access, and stack memory architectures
 for recurrent neural network language models. Our experiments on the Penn Treeb
ank and Wikitext-2 datasets show that stack-based memory architectures consisten
tly achieve the best performance in terms of held out perplexity. We also propos
e a generalization to existing continuous stack models (Joulin & Mikolov,2015; G
refenstette et al., 2015) to allow a variable number of pop operations more nat
urally that further improves performance. We further evaluate these language mod
els in terms of their ability to capture non-local syntactic dependencies on a s
ubject-verb agreement dataset (Linzen et al., 2016) and establish new state of

the art results using memory augmented language models. Our results demonstrate the value of stack-structured memory for explaining the distribution of words in natural language, in line with linguistic theories claiming a context-free backbone for natural language.
****
Wavelet Pooling for Convolutional Neural Networks
Travis Williams,Robert Li
Convolutional Neural Networks continuously advance the progress of 2D and 3D image and object classification. The steadfast usage of this algorithm requires constant evaluation and upgrading of foundational concepts to maintain progress. Network regularization techniques typically focus on convolutional layer operations, while leaving pooling layer operations without suitable options. We introduce Wavelet Pooling as another alternative to traditional neighborhood pooling. This method decomposes features into a second level decomposition, and discards the first-level subbands to reduce feature dimensions. This method addresses the overfitting problem encountered by max pooling, while reducing features in a more structurally compact manner than pooling via neighborhood regions. Experimental results on four benchmark classification datasets demonstrate our proposed method outperforms or performs comparatively with methods like max, mean, mixed, and stochastic pooling.
****
Deep Learning and Quantum Entanglement: Fundamental Connections with Implications to Network Design
Yoav Levine,David Yakira,Nadav Cohen,Amnon Shashua
Formal understanding of the inductive bias behind deep convolutional networks, i.e. the relation between the network's architectural features and the functions it is able to model, is limited. In this work, we establish a fundamental connection between the fields of quantum physics and deep learning, and use it for obtaining novel theoretical observations regarding the inductive bias of convolutional networks. Specifically, we show a structural equivalence between the function realized by a convolutional arithmetic circuit (ConvAC) and a quantum many-body wave function, which facilitates the use of quantum entanglement measures as quantifiers of a deep network's expressive ability to model correlations. Furthermore, the construction of a deep ConvAC in terms of a quantum Tensor Network is enabled. This allows us to perform a graph-theoretic analysis of a convolutional network, tying its expressiveness to a min-cut in its underlying graph. We demonstrate a practical outcome in the form of a direct control over the inductive bias via the number of channels (width) of each layer. We empirically validate our findings on standard convolutional networks which involve ReLU activations and max pooling. The description of a deep convolutional network in well-defined graph-theoretic tools and the structural connection to quantum entanglement, are two interdisciplinary bridges that are brought forth by this work.
****
Visualizing the Loss Landscape of Neural Nets
Hao Li,Zheng Xu,Gavin Taylor,Tom Goldstein
Neural network training relies on our ability to find ````````"good" minimizers of highly non-convex loss functions. It is well known that certain network architecture designs (e.g., skip connections) produce loss functions that train easier, and well-chosen training parameters (batch size, learning rate, optimizer) produce minimizers that generalize better. However, the reasons for these differences, and their effect on the underlying loss landscape, is not well understood.

In this paper, we explore the structure of neural loss functions, and the effect of loss landscapes on generalization, using a range of visualization methods. First, we introduce a simple ``"filter normalization" method that helps us visualize loss function curvature, and make meaningful side-by-side comparisons between loss functions. Then, using a variety of visualizations, we explore how network architecture effects the loss landscape, and how training parameters affect the shape of minimizers.
****

Variational Inference of Disentangled Latent Concepts from Unlabeled Observation s
Abhishek Kumar,Prasanna Sattigeri,Avinash Balakrishnan
Disentangled representations, where the higher level data generative factors are reflected in disjoint latent dimensions, offer several benefits such as ease of deriving invariant representations, transferability to other tasks, interpretab ility, etc. We consider the problem of unsupervised learning of disentangled rep resentations from large pool of unlabeled observations, and propose a variationa l inference based approach to infer disentangled latent factors. We introduce a regularizer on the expectation of the approximate posterior over observed data t hat encourages the disentanglement. We also propose a new disentanglement metric which is better aligned with the qualitative disentanglement observed in the de coder's output. We empirically observe significant improvement over existing met hods in terms of both disentanglement and data likelihood (reconstruction qualit y).

****
Pixel Deconvolutional Networks
Hongyang Gao,Hao Yuan,Zhengyang Wang,Shuiwang Ji
Deconvolutional layers have been widely used in a variety of deep models for up-sampling, including encoder-decoder networks for semantic segmentation and deep generative models for unsupervised learning. One of the key limitations of deconvolutional operations is that they result in the so-called checkerboard problem. This is caused by the fact that no direct relationship exists among adjacent pixels on the output feature map. To address this problem, we propose the pixel deconvolutional layer (PixelDCL) to establish direct relationships among adjacent pixels on the up-sampled feature map. Our method is based on a fresh interpretation of the regular deconvolution operation. The resulting PixelDCL can be used to replace any deconvolutional layer in a plug-and-play manner without compromising the fully trainable capabilities of original models. The proposed PixelDCL may result in slight decrease in efficiency, but this can be overcome by an implementation trick. Experimental results on semantic segmentation demonstrate that PixelDCL can consider spatial features such as edges and shapes and yields more accurate segmentation outputs than deconvolutional layers. When used in image generation tasks, our PixelDCL can largely overcome the checkerboard problem suffered by regular deconvolution operations.
****
Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecastin g
Yaguang Li,Rose Yu,Cyrus Shahabi,Yan Liu
Spatiotemporal forecasting has various applications in neuroscience, climate and transportation domain. Traffic forecasting is one canonical example of such lea rning task. The task is challenging due to (1) complex spatial dependency on roa d networks, (2) non-linear temporal dynamics with changing road conditions and (3) inherent difficulty of long-term forecasting. To address these challenges, we propose to model the traffic flow as a diffusion process on a directed graph an d introduce Diffusion Convolutional Recurrent Neural Network (DCRNN), a deep lea rning framework for traffic forecasting that incorporates both spatial and tempo ral dependency in the traffic flow. Specifically, DCRNN captures the spatial dep endency using bidirectional random walks on the graph, and the temporal dependen cy using the encoder-decoder architecture with scheduled sampling. We evaluate t he framework on two real-world large-scale road network traffic datasets and obs erve consistent improvement of 12% - 15% over state-of-the-art baselines
****
LSH Softmax: Sub-Linear Learning and Inference of the Softmax Layer in Deep Arch itectures

Daniel Levy,Danlu Chan,Stefano Ermon
Log-linear models models are widely used in machine learning, and in particular are ubiquitous in deep learning architectures in the form of the softmax. While exact inference and learning of these requires linear time, it can be done approximately in sub-linear time with strong concentrations guarantees. In this work, we present LSH Softmax, a method to perform sub-linear learning and inference of the softmax layer in the deep learning setting. Our method relies on the popular Locality-Sensitive Hashing to build a well-concentrated gradient estimator, using nearest neighbors and uniform samples. We also present an inference scheme in sub-linear time for LSH Softmax using the Gumbel distribution. On language modeling, we show that Recurrent Neural Networks trained with LSH Softmax perform on-par with computing the exact softmax while requiring sub-linear computations.
****
i-RevNet: Deep Invertible Networks
Jörn-Henrik Jacobsen,Arnold W.M. Smeulders,Edouard Oyallon
It is widely believed that the success of deep convolutional networks is based on progressively discarding uninformative variability about the input with respect to the problem at hand. This is supported empirically by the difficulty of recovering images from their hidden representations, in most commonly used network architectures. In this paper we show via a one-to-one mapping that this loss of information is not a necessary condition to learn representations that generalize well on complicated problems, such as ImageNet. Via a cascade of homeomorphic layers, we build the i-RevNet, a network that can be fully inverted up to the final projection onto the classes, i.e. no information is discarded. Building an invertible architecture is difficult, for one, because the local inversion is ill-conditioned, we overcome this by providing an explicit inverse.
An analysis of i-RevNet’s learned representations suggests an alternative explanation for the success of deep networks by a progressive contraction and linear separation with depth. To shed light on the nature of the model learned by the i-RevNet we reconstruct linear interpolations between natural image representations.
****
TESLA: Task-wise Early Stopping and Loss Aggregation for Dynamic Neural Network Inference
Chun-Min Chang,Chia-Ching Lin,Hung-Yi Ou Yang,Chin-Laung Lei,Kuan-Ta Chen
For inference operations in deep neural networks on end devices, it is desirable to deploy a single pre-trained neural network model, which can dynamically scale across a computation range without comprising accuracy. To achieve this goal, Incomplete Dot Product (IDP) has been proposed to use only a subset of terms in dot products during forward propagation. However, there are some limitations, including noticeable performance degradation in operating regions with low computational costs, and essential performance limitations since IDP uses hand-crafted profile coefficients. In this paper, we extend IDP by proposing new training algorithms involving a single profile, which may be trainable or pre-determined, to significantly improve the overall performance, especially in operating regions with low computational costs. Specifically, we propose the Task-wise Early Stopping and Loss Aggregation (TESLA) algorithm, which is showed in our 3-layer multilayer perceptron on MNIST that outperforms the original IDP by 32\% when only 10\% of dot products terms are used and achieves 94.7\% accuracy on average. By introducing trainable profile coefficients, TESLA further improves the accuracy to 95.5\% without specifying coefficients in advance. Besides, TESLA is applied to the VGG-16 model, which achieves 80\% accuracy using only 20\% of dot product terms on CIFAR-10 and also keeps 60\% accuracy using only 30\% of dot product terms on CIFAR-100, but the original IDP performs like a random guess in these two datasets at such low computation costs. Finally, we visualize the learned representations at different dot product percentages by class activation map and show that, by applying TESLA, the learned representations can adapt over a wide range of operation regions.
****
Unsupervised Cipher Cracking Using Discrete GANs

Aidan N. Gomez,Sicong Huang,Ivan Zhang,Bryan M. Li,Muhammad Osama,Lukasz Kaiser
This work details CipherGAN, an architecture inspired by CycleGAN used for infer
ring the underlying cipher mapping given banks of unpaired ciphertext and plaint
ext. We demonstrate that CipherGAN is capable of cracking language data encipher
ed using shift and Vigenere ciphers to a high degree of fidelity and for vocabul
aries much larger than previously achieved. We present how CycleGAN can be made
compatible with discrete data and train in a stable way. We then prove that the
technique used in CipherGAN avoids the common problem of uninformative discrimin
ation associated with GANs applied to discrete data.

****
Learning Parametric Closed-Loop Policies for Markov Potential Games
Sergio Valcarcel Macua,Javier Zazo,Santiago Zazo
Multiagent systems where the agents interact among themselves and with an stocha
stic environment can be formalized as stochastic games. We study a subclass of t
hese games, named Markov potential games (MPGs), that appear often in economic a
nd engineering applications when the agents share some common resource. We consi
der MPGs with continuous state-action variables, coupled constraints and nonconv
ex rewards. Previous analysis followed a variational approach that is only valid
 for very simple cases (convex rewards, invertible dynamics, and no coupled cons
traints); or considered deterministic dynamics and provided open-loop (OL) analy
sis, studying strategies that consist in predefined action sequences, which are
not optimal for stochastic environments. We present a closed-loop (CL) analysis
for MPGs and consider parametric policies that depend on the current state and w
here agents adapt to stochastic transitions. We provide easily verifiable, suffi
cient and necessary conditions for a stochastic game to be an MPG, even for comp
lex parametric functions (e.g., deep neural networks); and show that a closed-lo
op Nash equilibrium (NE) can be found (or at least approximated) by solving a re
lated optimal control problem (OCP). This is useful since solving an OCP---which
 is a single-objective problem---is usually much simpler than solving the origin
al set of coupled OCPs that form the game---which is a multiobjective control pr
oblem. This is a considerable improvement over the previously standard approach
for the CL analysis of MPGs, which gives no approximate solution if no NE belong
s to the chosen parametric family, and which is practical only for simple parame
tric forms. We illustrate the theoretical contributions with an example by apply
ing our approach to a noncooperative communications engineering game. We then so
lve the game with a deep reinforcement learning algorithm that learns policies t
hat closely approximates an exact variational NE of the game.
****
Unbiasing Truncated Backpropagation Through Time
Corentin Tallec,Yann Ollivier
\emph{Truncated Backpropagation Through Time} (truncated BPTT, \cite{jaeger2002t
utorial}) is a widespread method for learning recurrent computational graphs. Tr
uncated BPTT keeps the computational benefits of \emph{Backpropagation Through T
ime} (BPTT \cite{werbos:bptt}) while relieving the need for a complete backtrack
 through the whole data sequence at every step. However, truncation favors shor
t-term dependencies: the gradient estimate of truncated BPTT is biased, so that
it does not benefit from the convergence guarantees from stochastic gradient the
ory. We introduce \emph{Anticipated Reweighted Truncated Backpropagation} (ARTBP
), an algorithm that keeps the computational benefits of truncated BPTT, while p
roviding unbiasedness. ARTBP works by using variable truncation lengths together
 with carefully chosen compensation factors in the backpropagation equation. We
check the viability of ARTBP on two tasks. First, a simple synthetic task where
careful balancing of temporal dependencies at different scales is needed: trunca
ted BPTT displays unreliable performance, and in worst case scenarios, divergenc
e, while ARTBP converges reliably. Second, on Penn Treebank character-level lang
uage modelling \cite{ptb_proc}, ARTBP slightly outperforms truncated BPTT.

****
Post-training for Deep Learning

Thomas Moreau,Julien Audiffren
One of the main challenges of deep learning methods is the choice of an appropriate training strategy. In particular, additional steps, such as unsupervised pre-training, have been shown to greatly improve the performances of deep structures. In this article, we propose an extra training step, called post-training, which only optimizes the last layer of the network. We show that this procedure can be analyzed in the context of kernel theory, with the first layers computing an embedding of the data and the last layer a statistical model to solve the task based on this embedding. This step makes sure that the embedding, or representation, of the data is used in the best possible way for the considered task. This idea is then tested on multiple architectures with various data sets, showing that it consistently provides a boost in performance.
****
Transfer Learning on Manifolds via Learned Transport Operators
Marissa Connor,Christopher Rozell
Within-class variation in a high-dimensional dataset can be modeled as being on a low-dimensional manifold due to the constraints of the physical processes producing that variation (e.g., translation, illumination, etc.). We desire a method for learning a representation of the manifolds induced by identity-preserving transformations that can be used to increase robustness, reduce the training burden, and encourage interpretability in machine learning tasks. In particular, what is needed is a representation of the transformation manifold that can robustly capture the shape of the manifold from the input data, generate new points on the manifold, and extend transformations outside of the training domain without significantly increasing the error. Previous work has proposed algorithms to efficiently learn analytic operators (called transport operators) that define the process of transporting one data point on a manifold to another. The main contribution of this paper is to define two transfer learning methods that use this generative manifold representation to learn natural transformations and incorporate them into new data. The first method uses this representation in a novel randomized approach to transfer learning that employs the learned generative model to map out unseen regions of the data space. These results are shown through demonstrations of transfer learning in a data augmentation task for few-shot image classification. The second method use of transport operators for injecting specific transformations into new data examples which allows for realistic image animation and informed data augmentation. These results are shown on stylized constructions using the classic swiss roll data structure and in demonstrations of transfer learning in a data augmentation task for few-shot image classification. We also propose the use of transport operators for injecting transformations into new data examples which allows for realistic image animation.
****
PARAMETRIZED DEEP Q-NETWORKS LEARNING: PLAYING ONLINE BATTLE ARENA WITH DISCRETE-CONTINUOUS HYBRID ACTION SPACE
Jiechao Xiong,Qing Wang,Zhuoran Yang,Peng Sun,Yang Zheng,Lei Han,Haobo Fu,Xiangru Lian,Carson Eisenach,Haichuan Yang,Emmanuel Ekwedike,Bei Peng,Haoyue Gao,Tong Zhang,Ji Liu,Han Liu
Most existing deep reinforcement learning (DRL) frameworks consider action spaces that are either
discrete or continuous space. Motivated by the project of design Game AI for King of Glory
(KOG), one the world’s most popular mobile game, we consider the scenario with the discrete-continuous
hybrid action space. To directly apply existing DLR frameworks, existing approaches
either approximate the hybrid space by a discrete set or relaxing it into a continuous set, which is
usually less efficient and robust. In this paper, we propose a parametrized deep Q-network (P-DQN)
for the hybrid action space without approximation or relaxation. Our algorithm combines DQN and

DDPG and can be viewed as an extension of the DQN to hybrid actions. The empiric al study on the
game KOG validates the efficiency and effectiveness of our method.
****
Better Generalization by Efficient Trust Region Method
Xuanqing Liu,Jason D. Lee,Cho-Jui Hsieh
In this paper, we develop a trust region method for training deep neural network s. At each iteration, trust region method computes the search direction by solvi ng a non-convex subproblem. Solving this subproblem is non-trivial---existing me thods have only sub-linear convergence rate. In the first part, we show that a s imple modification of gradient descent algorithm can converge to a global minimi zer of the subproblem with an asymptotic linear convergence rate. Moreover, our method only requires Hessian-vector products, which can be computed efficiently by back-propagation in neural networks. In the second part, we apply our algorit hm to train large-scale convolutional neural networks, such as VGG and MobileNet s. Although trust region method is about 3 times slower than SGD in terms of run ning time, we observe it finds a model that has lower generalization (test) erro r than SGD, and this difference is even more significant in large batch training .
We conduct several interesting experiments to support our conjecture that the tr ust region method can avoid sharp local minimas.
****
Contextual Explanation Networks
Maruan Al-Shedivat,Avinava Dubey,Eric P. Xing
We introduce contextual explanation networks (CENs)---a class of models that lea rn to predict by generating and leveraging intermediate explanations. CENs are d eep networks that generate parameters for context-specific probabilistic graphic al models which are further used for prediction and play the role of explanation s. Contrary to the existing post-hoc model-explanation tools, CENs learn to pred ict and to explain jointly. Our approach offers two major advantages: (i) for ea ch prediction, valid instance-specific explanations are generated with no comput ational overhead and (ii) prediction via explanation acts as a regularization an d boosts performance in low-resource settings. We prove that local approximation s to the decision boundary of our networks are consistent with the generated exp lanations. Our results on image and text classification and survival analysis ta sks demonstrate that CENs are competitive with the state-of-the-art while offeri ng additional insights behind each prediction, valuable for decision support.
****
Temporal Difference Models: Model-Free Deep RL for Model-Based Control
Vitchyr Pong*,Shixiang Gu*,Murtaza Dalal,Sergey Levine
Model-free reinforcement learning (RL) has been proven to be a powerful, general tool for learning complex behaviors. However, its sample efficiency is often im practically large for solving challenging real-world problems, even for off-poli cy algorithms such as Q-learning. A limiting factor in classic model-free RL is that the learning signal consists only of scalar rewards, ignoring much of the r ich information contained in state transition tuples. Model-based RL uses this i nformation, by training a predictive model, but often does not achieve the same asymptotic performance as model-free RL due to model bias. We introduce temporal difference models (TDMs), a family of goal-conditioned value functions that can be trained with model-free learning and used for model-based control. TDMs comb ine the benefits of model-free and model-based RL: they leverage the rich inform ation in state transitions to learn very efficiently, while still attaining asym ptotic performance that exceeds that of direct model-based RL methods. Our exper imental results show that, on a range of continuous control tasks, TDMs provide a substantial improvement in efficiency compared to state-of-the-art model-based and model-free methods.
****
ENRICHMENT OF FEATURES FOR CLASSIFICATION USING AN OPTIMIZED LINEAR/NON-LINEAR C OMBINATION OF INPUT FEATURES
Mehran Taghipour-Gorjikolaie,Seyyed Mohammad Razavi,Javad Sadri

Automatic classification of objects is one of the most important tasks in engine ering
and data mining applications. Although using more complex and advanced
classifiers can help to improve the accuracy of classification systems, it can b e
done by analyzing data sets and their features for a particular problem. Feature
combination is the one which can improve the quality of the features. In this pa per,
a structure similar to Feed-Forward Neural Network (FFNN) is used to generate an
optimized linear or non-linear combination of features for classification. Genet ic
Algorithm (GA) is applied to update weights and biases. Since nature of data set s
and their features impact on the effectiveness of combination and classification
system, linear and non-linear activation functions (or transfer function) are us ed
to achieve more reliable system. Experiments of several UCI data sets and using
minimum distance classifier as a simple classifier indicate that proposed linear and
non-linear intelligent FFNN-based feature combination can present more reliable
and promising results. By using such a feature combination method, there is no
need to use more powerful and complex classifier anymore.
****

Model compression via distillation and quantization

Antonio Polino,Razvan Pascanu,Dan Alistarh

Deep neural networks (DNNs) continue to make significant advances, solving tasks from image classification to translation or reinforcement learning. One aspect of the field receiving considerable attention is efficiently executing deep mode ls in resource-constrained environments, such as mobile or embedded devices. Thi s paper focuses on this problem, and proposes two new compression methods, which jointly leverage weight quantization and distillation of larger teacher network s into smaller student networks. The first method we propose is called quantized distillation and leverages distillation during the training process, by incorpo rating distillation loss, expressed with respect to the teacher, into the traini ng of a student network whose weights are quantized to a limited set of levels. The second method, differentiable quantization, optimizes the location of quant ization points through stochastic gradient descent, to better fit the behavior o f the teacher model. We validate both methods through experiments on convolutio nal and recurrent architectures. We show that quantized shallow students can rea ch similar accuracy levels to full-precision teacher models, while providing ord er of magnitude compression, and inference speedup that is linear in the depth r eduction. In sum, our results enable DNNs for resource-constrained environments to leverage architecture and accuracy advances developed on more powerful device s.

****

Correcting Nuisance Variation using Wasserstein Distance

Gil Tabak,Minjie Fan,Samuel J. Yang,Stephan Hoyer,Geoff Davis

Profiling cellular phenotypes from microscopic imaging can provide meaningful bi ological information resulting from various factors affecting the cells. One mot ivating application is drug development: morphological cell features can be capt ured from images, from which similarities between different drugs applied at dif ferent dosages can be quantified. The general approach is to find a function map ping the images to an embedding space of manageable dimensionality whose geometr y captures relevant features of the input images. An important known issue for s uch methods is separating relevant biological signal from nuisance variation. Fo r example, the embedding vectors tend to be more correlated for cells that were cultured and imaged during the same week than for cells from a different week, d espite having identical drug compounds applied in both cases. In this case, the particular batch a set of experiments were conducted in constitutes the domain o

f the data; an ideal set of image embeddings should contain only the relevant biological information (e.g. drug effects). We develop a general framework for adjusting the image embeddings in order to `forget' domain-specific information while preserving relevant biological information. To do this, we minimize a loss function based on distances between marginal distributions (such as the Wasserstein distance) of embeddings across domains for each replicated treatment. For the dataset presented, the replicated treatment is the negative control. We find that for our transformed embeddings (1) the underlying geometric structure is not only preserved but the embeddings also carry improved biological signal (2) less domain-specific information is present.
****
TCAV: Relative concept importance testing with Linear Concept Activation Vectors
Been Kim,Justin Gilmer,Martin Wattenberg,Fernanda Viégas
Despite neural network’s high performance, the lack of interpretability has been the main bottleneck for its safe usage in practice. In domains with high stakes (e.g., medical diagnosis), gaining insights into the network is critical for gaining trust and being adopted. One of the ways to improve interpretability of a NN is to explain the importance of a particular concept (e.g., gender) in prediction. This is useful for explaining reasoning behind the networks’ predictions, and for revealing any biases the network may have. This work aims to provide quantitative answers to \textit{the relative importance of concepts of interest} via concept activation vectors (CAV). In particular, this framework enables non-machine learning experts to express concepts of interests and test hypotheses using examples (e.g., a set of pictures that illustrate the concept). We show that CAV can be learned given a relatively small set of examples. Testing with CAV, for example, can answer whether a particular concept (e.g., gender) is more important in predicting a given class (e.g., doctor) than other set of concepts. Interpreting with CAV does not require any retraining or modification of the network. We show that many levels of meaningful concepts are learned (e.g., color, texture, objects, a person’s occupation), and we present CAV’s \textit{empirical deepdream} — where we maximize an activation using a set of example pictures. We show how various insights can be gained from the relative importance testing with CAV.
****
Jointly Learning Sentence Embeddings and Syntax with Unsupervised Tree-LSTMs
Jean Maillard,Stephen Clark,Dani Yogatama
We introduce a neural network that represents sentences by composing their words according to induced binary parse trees. We use Tree-LSTM as our composition function, applied along a tree structure found by a fully differentiable natural language chart parser. Our model simultaneously optimises both the composition function and the parser, thus eliminating the need for externally-provided parse trees which are normally required for Tree-LSTM. It can therefore be seen as a tree-based RNN that is unsupervised with respect to the parse trees. As it is fully differentiable, our model is easily trained with an off-the-shelf gradient descent method and backpropagation. We demonstrate that it achieves better performance compared to various supervised Tree-LSTM architectures on a textual entailment task and a reverse dictionary task. Finally, we show how performance can be improved with an attention mechanism which fully exploits the parse chart, by attending over all possible subspans of the sentence.
****
Auto-Conditioned Recurrent Networks for Extended Complex Human Motion Synthesis
Yi Zhou,Zimo Li,Shuangjiu Xiao,Chong He,Zeng Huang,Hao Li
We present a real-time method for synthesizing highly complex human motions using a novel training regime we call the auto-conditioned Recurrent Neural Network (acRNN). Recently, researchers have attempted to synthesize new motion by using autoregressive techniques, but existing methods tend to freeze or diverge after a couple of seconds due to an accumulation of errors that are fed back into the network. Furthermore, such methods have only been shown to be reliable for relatively simple human motions, such as walking or running. In contrast, our approach can synthesize arbitrary motions with highly complex styles, including dances

or martial arts in addition to locomotion. The acRNN is able to accomplish this by explicitly accommodating for autoregressive noise accumulation during training. Our work is the first to our knowledge that demonstrates the ability to generate over 18,000 continuous frames (300 seconds) of new complex human motion w.r.t. different styles.
****
On the Expressive Power of Overlapping Architectures of Deep Learning
Or Sharir,Amnon Shashua
Expressive efficiency refers to the relation between two architectures A and B, whereby any function realized by B could be replicated by A, but there exists functions realized by A, which cannot be replicated by B unless its size grows significantly larger. For example, it is known that deep networks are exponentially efficient with respect to shallow networks, in the sense that a shallow network must grow exponentially large in order to approximate the functions represented by a deep network of polynomial size. In this work, we extend the study of expressive efficiency to the attribute of network connectivity and in particular to the effect of "overlaps" in the convolutional process, i.e., when the stride of the convolution is smaller than its filter size (receptive field).
To theoretically analyze this aspect of network's design, we focus on a well-established surrogate for ConvNets called Convolutional Arithmetic Circuits (ConvACs), and then demonstrate empirically that our results hold for standard ConvNets as well. Specifically, our analysis shows that having overlapping local receptive fields, and more broadly denser connectivity, results in an exponential increase in the expressive capacity of neural networks. Moreover, while denser connectivity can increase the expressive capacity, we show that the most common types of modern architectures already exhibit exponential increase in expressivity, without relying on fully-connected layers.
****
Regularizing and Optimizing LSTM Language Models
Stephen Merity,Nitish Shirish Keskar,Richard Socher
In this paper, we consider the specific problem of word-level language modeling and investigate strategies for regularizing and optimizing LSTM-based models. We propose the weight-dropped LSTM, which uses DropConnect on hidden-to-hidden weights, as a form of recurrent regularization. Further, we introduce NT-ASGD, a non-monotonically triggered (NT) variant of the averaged stochastic gradient method (ASGD), wherein the averaging trigger is determined using a NT condition as opposed to being tuned by the user. Using these and other regularization strategies, our ASGD Weight-Dropped LSTM (AWD-LSTM) achieves state-of-the-art word level perplexities on two data sets: 57.3 on Penn Treebank and 65.8 on WikiText-2. In exploring the effectiveness of a neural cache in conjunction with our proposed model, we achieve an even lower state-of-the-art perplexity of 52.8 on Penn Treebank and 52.0 on WikiText-2. We also explore the viability of the proposed regularization and optimization strategies in the context of the quasi-recurrent neural network (QRNN) and demonstrate comparable performance to the AWD-LSTM counterpart. The code for reproducing the results is open sourced and is available at https://github.com/salesforce/awd-lstm-lm.
****
Piecewise Linear Neural Networks verification: A comparative study
Rudy Bunel,Ilker Turkaslan,Philip H.S. Torr,Pushmeet Kohli,M. Pawan Kumar
The success of Deep Learning and its potential use in many important safety-critical applications has motivated research on formal verification of Neural Net-
work (NN) models. Despite the reputation of learned NN models to behave as black boxes and theoretical hardness results of the problem of proving their prop-
erties, researchers have been successful in verifying some classes of models by exploiting their piecewise linear structure. Unfortunately, most of these works test their algorithms on their own models and do not offer any comparison with other approaches. As a result, the advantages and downsides of the different al-
gorithms are not well understood. Motivated by the need of accelerating progress

in this very important area, we investigate the trade-offs of a number of differ
ent
approaches based on Mixed Integer Programming, Satisfiability Modulo Theory,
as well as a novel method based on the Branch-and-Bound framework. We also
propose a new data set of benchmarks, in addition to a collection of previously
released testcases that can be used to compare existing methods. Our analysis no
t
only allowed a comparison to be made between different strategies, the compar-
ision of results from different solvers also revealed implementation bugs in pub
-
lished methods. We expect that the availability of our benchmark and the analysi
s
of the different approaches will allow researchers to invent and evaluate promis
ing
approaches for making progress on this important topic.
****
Memorization Precedes Generation: Learning Unsupervised GANs with Memory Network
s
Youngjin Kim,Minjung Kim,Gunhee Kim
We propose an approach to address two issues that commonly occur during training
 of unsupervised GANs. First, since GANs use only a continuous latent distributi
on to embed multiple classes or clusters of data, they often do not correctly ha
ndle the structural discontinuity between disparate classes in a latent space. S
econd, discriminators of GANs easily forget about past generated samples by gene
rators, incurring instability during adversarial training. We argue that these t
wo infamous problems of unsupervised GAN training can be largely alleviated by a
 learnable memory network to which both generators and discriminators can access
. Generators can effectively learn representation of training samples to underst
and underlying cluster distributions of data, which ease the structure discontin
uity problem. At the same time, discriminators can better memorize clusters of p
reviously generated samples, which mitigate the forgetting problem. We propose a
 novel end-to-end GAN model named memoryGAN, which involves a memory network tha
t is unsupervisedly trainable and integrable to many existing GAN models. With e
valuations on multiple datasets such as Fashion-MNIST, CelebA, CIFAR10, and Chai
rs, we show that our model is probabilistically interpretable, and generates rea
listic image samples of high visual fidelity. The memoryGAN also achieves the st
ate-of-the-art inception scores over unsupervised GAN models on the CIFAR10 data
set, without any optimization tricks and weaker divergences.
****
An image representation based convolutional network for DNA classification
Bojian Yin,Marleen Balvert,Davide Zambrano,Alexander Schoenhuth,Sander Bohte
The folding structure of the DNA molecule combined with helper molecules, also r
eferred to as the chromatin, is highly relevant for the functional properties of
 DNA. The chromatin structure is largely determined by the underlying primary DN
A sequence, though the interaction is not yet fully understood. In this paper we
 develop a convolutional neural network that takes an image-representation of pr
imary DNA sequence as its input, and predicts key determinants of chromatin stru
cture. The method is developed such that it is capable of detecting interactions
 between distal elements in the DNA sequence, which are known to be highly relev
ant. Our experiments show that the method outperforms several existing methods b
oth in terms of prediction accuracy and training time.
****
Iterative Deep Compression : Compressing Deep Networks for Classification and Se
mantic Segmentation
Sugandha Doda,Vitor Fortes Rey,Dr. Nadereh Hatami,Prof. Dr. Paul Lukowicz
Machine learning and in particular deep learning approaches have outperformed ma
ny traditional techniques in accomplishing complex tasks such as
image classfication, natural language processing or speech recognition. Most of
the state-of-the art deep networks have complex architecture and use a vast numb
er of parameters to reach this superior performance. Though these networks use a

large number of learnable parameters, those parameters present significant redundancy. Therefore, it is possible to compress the network without much affecting its accuracy by eliminating those redundant and unimportant parameters.
In this work, we propose a three stage compression pipeline, which consists of pruning, weight sharing and quantization to compress deep neural networks.
Our novel pruning technique combines magnitude based ones with dense sparse dense ideas and iteratively finds for each layer its achievable sparsity instead of selecting a single threshold for the whole network.
Unlike previous works, where compression is only applied on networks performing classification, we evaluate and perform compression on networks for classification as well as semantic segmentation, which is greatly useful for understanding scenes in autonomous driving.
We tested our method on LeNet-5 and FCNs, performing classification and semantic segmentation, respectively. With LeNet-5 on MNIST, pruning reduces the number of parameters by 15.3 times and storage requirement from 1.7 MB to 0.006 MB with accuracy loss of 0.03%. With FCN8 on Cityscapes, we decrease the number of parameters by 8 times and reduce the storage requirement from 537.47 MB to 18.23 MB with class-wise intersection-over-union (IoU) loss of 4.93% on the validation data.
****
Simple and efficient architecture search for Convolutional Neural Networks
Thomas Elsken,Jan Hendrik Metzen,Frank Hutter
Neural networks have recently had a lot of success for many tasks. However, neural
network architectures that perform well are still typically designed manually
by experts in a cumbersome trial-and-error process. We propose a new method
to automatically search for well-performing CNN architectures based on a simple
hill climbing procedure whose operators apply network morphisms, followed
by short optimization runs by cosine annealing. Surprisingly, this simple method
yields competitive results, despite only requiring resources in the same order of
magnitude as training a single network. E.g., on CIFAR-10, our method designs
and trains networks with an error rate below 6% in only 12 hours on a single GPU;
training for one day reduces this error further, to almost 5%.
****
Stochastic Training of Graph Convolutional Networks
Jianfei Chen,Jun Zhu
Graph convolutional networks (GCNs) are powerful deep neural networks for graph-structured data. However, GCN computes nodes' representation recursively from their neighbors, making the receptive field size grow exponentially with the number of layers. Previous attempts on reducing the receptive field size by subsampling neighbors do not have any convergence guarantee, and their receptive field size per node is still in the order of hundreds. In this paper, we develop a preprocessing strategy and two control variate based algorithms to further reduce the receptive field size. Our algorithms are guaranteed to converge to GCN's local optimum regardless of the neighbor sampling size. Empirical results show that our algorithms have a similar convergence speed per epoch with the exact algorithm even using only two neighbors per node. The time consumption of our algorithm on the Reddit dataset is only one fifth of previous neighbor sampling algorithms.
****
Discrete-Valued Neural Networks Using Variational Inference
Wolfgang Roth,Franz Pernkopf
The increasing demand for neural networks (NNs) being employed on embedded devices has led to plenty of research investigating methods for training low precision NNs. While most methods involve a quantization step, we propose a principled Bayesian approach where we first infer a distribution over a discrete weight space from which we subsequently derive hardware-friendly low precision NNs. To this end, we introduce a probabilistic forward pass to approximate the intractable v

ariational objective that allows us to optimize over discrete-valued weight distributions for NNs with sign activation functions. In our experiments, we show that our model achieves state of the art performance on several real world data sets. In addition, the resulting models exhibit a substantial amount of sparsity that can be utilized to further reduce the computational costs for inference.
****

Revisiting Knowledge Base Embedding as Tensor Decomposition

Jiezhong Qiu,Hao Ma,Yuxiao Dong,Kuansan Wang,Jie Tang

We study the problem of knowledge base (KB) embedding, which is usually addressed through two frameworks---neural KB embedding and tensor decomposition. In this work, we theoretically analyze the neural embedding framework and subsequently connect it with tensor based embedding. Specifically, we show that in neural KB embedding the two commonly adopted optimization solutions---margin-based and negative sampling losses---are closely related to each other. We also reach the closed-form tensor that is implicitly approximated by popular neural KB approaches, revealing the underlying connection between neural and tensor based KB embedding models. Grounded in the theoretical results, we further present a tensor decomposition based framework KBTD to directly approximate the derived closed form tensor. Under this framework, the neural KB embedding models, such as NTN, TransE, Bilinear, and DISTMULT, are unified into a general tensor optimization architecture. Finally, we conduct experiments on the link prediction task in WordNet and Freebase, empirically demonstrating the effectiveness of the KBTD framework.

****

Autostacker: an Automatic Evolutionary Hierarchical Machine Learning System

Boyuan Chen,Warren Mo,Ishanu Chattopadhyay,Hod Lipson

This work provides an automatic machine learning (AutoML) modelling architecture called Autostacker. Autostacker improves the prediction accuracy of machine learning baselines by utilizing an innovative hierarchical stacking architecture and an efficient parameter search algorithm. Neither prior domain knowledge about the data nor feature preprocessing is needed. We significantly reduce the time of AutoML with a naturally inspired algorithm - Parallel Hill Climbing (PHC). By parallelizing PHC, Autostacker can provide candidate pipelines with sufficient prediction accuracy within a short amount of time. These pipelines can be used as is or as a starting point for human experts to build on. By focusing on the modelling process, Autostacker breaks the tradition of following fixed order pipelines by exploring not only single model pipeline but also innovative combinations and structures. As we will show in the experiment section, Autostacker achieves significantly better performance both in terms of test accuracy and time cost comparing with human initial trials and recent popular AutoML system.
****

LEARNING SEMANTIC WORD RESPRESENTATIONS VIA TENSOR FACTORIZATION

Eric Bailey,Charles Meyer,Shuchin Aeron

Many state-of-the-art word embedding techniques involve factorization of a cooccurrence
based matrix. We aim to extend this approach by studying word embedding
techniques that involve factorization of co-occurrence based tensors (N-way arrays). We present two new word embedding techniques based on tensor
factorization and show that they outperform common methods on several semantic
NLP tasks when given the same data. To train one of the embeddings, we present
a new joint tensor factorization problem and an approach for solving it. Furthermore,
we modify the performance metrics for the Outlier Detection Camacho-Collados & Navigli (2016) task to measure the quality of higher-order relationships
that a word embedding captures. Our tensor-based methods significantly
outperform existing methods at this task when using our new metric. Finally, we
demonstrate that vectors in our embeddings can be composed multiplicatively to
create different vector representations for each meaning of a polysemous word.
We show that this property stems from the higher order information that the vect

ors
contain, and thus is unique to our tensor based embeddings.
****
Reinforcement Learning on Web Interfaces using Workflow-Guided Exploration
Evan Zheran Liu,Kelvin Guu,Panupong Pasupat,Tianlin Shi,Percy Liang
Reinforcement learning (RL) agents improve through trial-and-error, but when reward is sparse and the agent cannot discover successful action sequences, learning stagnates. This has been a notable problem in training deep RL agents to perform web-based tasks, such as booking flights or replying to emails, where a single mistake can ruin the entire sequence of actions. A common remedy is to "warm-start" the agent by pre-training it to mimic expert demonstrations, but this is prone to overfitting. Instead, we propose to constrain exploration using demonstrations. From each demonstration, we induce high-level "workflows" which constrain the allowable actions at each time step to be similar to those in the demonstration (e.g., "Step 1: click on a textbox; Step 2: enter some text"). Our exploration policy then learns to identify successful workflows and samples actions that satisfy these workflows. Workflows prune out bad exploration directions and accelerate the agent’s ability to discover rewards. We use our approach to train a novel neural policy designed to handle the semi-structured nature of websites, and evaluate on a suite of web tasks, including the recent World of Bits benchmark. We achieve new state-of-the-art results, and show that workflow-guided exploration improves sample efficiency over behavioral cloning by more than 100x.
****
How do deep convolutional neural networks learn from raw audio waveforms?
Yuan Gong,Christian Poellabauer
Prior work on speech and audio processing has demonstrated the ability to obtain excellent performance when learning directly from raw audio waveforms using convolutional neural networks (CNNs). However, the exact inner workings of a CNN remain unclear, which hinders further developments and improvements into this direction. In this paper, we theoretically analyze and explain how deep CNNs learn from raw audio waveforms and identify potential limitations of existing network structures. Based on this analysis, we further propose a new network architecture (called SimpleNet), which offers a very simple but concise structure and high model interpretability.
****
Thermometer Encoding: One Hot Way To Resist Adversarial Examples
Jacob Buckman,Aurko Roy,Colin Raffel,Ian Goodfellow
It is well known that it is possible to construct "adversarial examples"
for neural networks: inputs which are misclassified by the network
yet indistinguishable from true data. We propose a simple
modification to standard neural network architectures, thermometer
encoding, which significantly increases the robustness of the network to
adversarial examples. We demonstrate this robustness with experiments
on the MNIST, CIFAR-10, CIFAR-100, and SVHN datasets, and show that
models with thermometer-encoded inputs consistently have higher accuracy
on adversarial examples, without decreasing generalization.
State-of-the-art accuracy under the strongest known white-box attack was
increased from 93.20% to 94.30% on MNIST and 50.00% to 79.16% on CIFAR-10.
We explore the properties of these networks, providing evidence
that thermometer encodings help neural networks to
find more-non-linear decision boundaries.
****
Characterizing Sparse Connectivity Patterns in Neural Networks
Sourya Dey,Kuan-Wen Huang,Peter A. Beerel,Keith M. Chugg
We propose a novel way of reducing the number of parameters in the storage-hungry fully connected layers of a neural network by using pre-defined sparsity, where the majority of connections are absent prior to starting training. Our results indicate that convolutional neural networks can operate without any loss of accuracy at less than 0.5% classification layer connection density, or less than 5% overall network connection density. We also investigate the effects of pre-defi

ning the sparsity of networks with only fully connected layers. Based on our spa rsifying technique, we introduce the `scatter' metric to characterize the qualit y of a particular connection pattern. As proof of concept, we show results on CI FAR, MNIST and a new dataset on classifying Morse code symbols, which highlights some interesting trends and limits of sparse connection patterns.
****

Learning Document Embeddings With CNNs
Shunan Zhao,Chundi Lui,Maksims Volkovs
This paper proposes a new model for document embedding. Existing approaches eith er require complex inference or use recurrent neural networks that are difficult to parallelize. We take a different route and use recent advances in language m odeling to develop a convolutional neural network embedding model. This allows u s to train deeper architectures that are fully parallelizable. Stacking layers t ogether increases the receptive filed allowing each successive layer to model in creasingly longer range semantic dependences within the document. Empirically we demonstrate superior results on two publicly available benchmarks. Full code wi ll be released with the final version of this paper.
****

Fixing Weight Decay Regularization in Adam
Ilya Loshchilov,Frank Hutter
We note that common implementations of adaptive gradient algorithms, such as Ada m, limit the potential benefit of weight decay regularization, because the weigh ts do not decay multiplicatively (as would be expected for standard weight decay) but by an additive constant factor.
We propose a simple way to resolve this issue by decoupling weight decay and the optimization steps taken w.r.t. the loss function. We provide empirical evidenc e that our proposed modification (i)
decouples the optimal choice of weight decay factor from the setting of the lear ning rate for both standard SGD and Adam, and (ii) substantially improves Adam's generalization performance, allowing it to compete with SGD with momentum on im age classification datasets (on which it was previously typically outperformed b y the latter).
We also demonstrate that longer optimization runs require smaller weight decay v alues for optimal results and introduce a normalized variant of weight decay to reduce this dependence. Finally, we propose a version of Adam with warm restarts (AdamWR) that has strong anytime performance while achieving state-of-the-art r esults on CIFAR-10 and ImageNet32x32.
Our source code will become available after the review process.
****

Feature Incay for Representation Regularization
Yuhui Yuan,Kuiyuan Yang,Jianyuan Guo,Jingdong Wang,Chao Zhang
Softmax-based loss is widely used in deep learning for multi-class classificatio n, where each class is represented by a weight vector and each sample is represe nted as a feature vector. Different from traditional learning algorithms where f eatures are pre-defined and only weight vectors are tunable through training, fe ature vectors are also tunable as representation learning in deep learning. Thus we investigate how to improve the classification performance by better adjustin g the features. One main observation is that elongating the feature norm of both correctly-classified and mis-classified feature vectors improves learning: (1) increasing the feature norm of correctly-classified examples induce smaller trai ning loss; (2) increasing the feature norm of mis-classified examples can upweig ht the contribution from hard examples. Accordingly, we propose feature incay to regularize representation learning by encouraging larger feature norm. In contr ast to weight decay which shrinks the weight norm, feature incay is proposed to stretch the feature norm. Extensive empirical results on MNIST, CIFAR10, CIFAR10 0 and LFW demonstrate the effectiveness of feature incay.
****

On the State of the Art of Evaluation in Neural Language Models
Gábor Melis,Chris Dyer,Phil Blunsom
Ongoing innovations in recurrent neural network architectures have provided a st

eady influx of apparently state-of-the-art results on language modelling benchmarks. However, these have been evaluated using differing codebases and limited computational resources, which represent uncontrolled sources of experimental variation. We reevaluate several popular architectures and regularisation methods with large-scale automatic black-box hyperparameter tuning and arrive at the somewhat surprising conclusion that standard LSTM architectures, when properly regularised, outperform more recent models. We establish a new state of the art on the Penn Treebank and Wikitext-2 corpora, as well as strong baselines on the Hutter Prize dataset.

****
Sensitivity and Generalization in Neural Networks: an Empirical Study
Roman Novak,Yasaman Bahri,Daniel A. Abolafia,Jeffrey Pennington,Jascha Sohl-Dickstein
In practice it is often found that large over-parameterized neural networks generalize better than their smaller counterparts, an observation that appears to conflict with classical notions of function complexity, which typically favor smaller models. In this work, we investigate this tension between complexity and generalization through an extensive empirical exploration of two natural metrics of complexity related to sensitivity to input perturbations. Our experiments survey thousands of models with different architectures, optimizers, and other hyper-parameters, as well as four different image classification datasets.

We find that trained neural networks are more robust to input perturbations in the vicinity of the training data manifold, as measured by the input-output Jacobian of the network, and that this correlates well with generalization. We further establish that factors associated with poor generalization -- such as full-batch training or using random labels -- correspond to higher sensitivity, while factors associated with good generalization -- such as data augmentation and ReLU non-linearities -- give rise to more robust functions. Finally, we demonstrate how the input-output Jacobian norm can be predictive of generalization at the level of individual test points.
****
Simulating Action Dynamics with Neural Process Networks
Antoine Bosselut,Omer Levy,Ari Holtzman,Corin Ennis,Dieter Fox,Yejin Choi
Understanding procedural language requires anticipating the causal effects of actions, even when they are not explicitly stated. In this work, we introduce Neural Process Networks to understand procedural text through (neural) simulation of action dynamics.   Our model complements existing memory architectures with dynamic entity tracking by explicitly modeling actions as state transformers. The model updates the states of the entities by executing learned action operators. Empirical results demonstrate that our proposed model can reason about the unstated causal effects of actions, allowing it to provide more accurate contextual information for understanding and generating procedural text, all while offering more interpretable internal representations than existing alternatives.
****
CausalGAN: Learning Causal Implicit Generative Models with Adversarial Training
Murat Kocaoglu,Christopher Snyder,Alexandros G. Dimakis,Sriram Vishwanath
We introduce causal implicit generative models (CiGMs): models that allow sampling from not only the true observational but also the true interventional distributions. We show that adversarial training can be used to learn a CiGM, if the generator architecture is structured based on a given causal graph. We consider the application of conditional and interventional sampling of face images with binary feature labels, such as mustache, young. We preserve the dependency structure between the labels with a given causal graph. We devise a two-stage procedure for learning a CiGM over the labels and the image. First we train a CiGM over the binary labels using a Wasserstein GAN where the generator neural network is consistent with the causal graph between the labels. Later, we combine this with a conditional GAN to generate images conditioned on the binary labels. We propose two new conditional GAN architectures: CausalGAN and CausalBEGAN. We show that

the optimal generator of the CausalGAN, given the labels, samples from the imag e distributions conditioned on these labels. The conditional GAN combined with a trained CiGM for the labels is then a CiGM over the labels and the generated im age. We show that the proposed architectures can be used to sample from observat ional and interventional image distributions, even for interventions which do no t naturally occur in the dataset.
****

Learning Covariate-Specific Embeddings with Tensor Decompositions
Kevin Tian,Teng Zhang,James Zou
Word embedding is a useful approach to capture co-occurrence structures in a lar ge corpus of text. In addition to the text data itself, we often have additional covariates associated with individual documents in the corpus---e.g. the demogr aphic of the author, time and venue of publication, etc.---and we would like the embedding to naturally capture the information of the covariates. In this paper , we propose a new tensor decomposition model for word embeddings with covariate s. Our model jointly learns a \emph{base} embedding for all the words as well as a weighted diagonal transformation to model how each covariate modifies the bas e embedding. To obtain the specific embedding for a particular author or venue, for example, we can then simply multiply the base embedding by the transformatio n matrix associated with that time or venue. The main advantages of our approach is data efficiency and interpretability of the covariate transformation matrix. Our experiments demonstrate that our joint model learns substantially better em beddings conditioned on each covariate compared to the standard approach of lear ning a separate embedding for each covariate using only the relevant subset of d ata. Furthermore, our model encourages the embeddings to be ``topic-aligned'' in the sense that the dimensions have specific independent meanings. This allows o ur covariate-specific embeddings to be compared by topic, enabling downstream di fferential analysis. We empirically evaluate the benefits of our algorithm on se veral datasets, and demonstrate how it can be used to address many natural quest ions about the effects of covariates.
****

Espresso: Efficient Forward Propagation for Binary Deep Neural Networks
Fabrizio Pedersoli,George Tzanetakis,Andrea Tagliasacchi
There are many applications scenarios for which the computational performance and memory footprint of the prediction phase of Deep Neural Networks (DNNs) need to be optimized. Binary Deep Neural Networks (BDNNs) have been shown to be an effective way of achieving this objective. In this paper, we show how Convolutional Neural Networks (CNNs) can be implemented using binary representations. Espresso is a compact, yet powerful library written in C/CUDA that features all the functionalities required for the forward propagation of CNNs, in a binary file less than 400KB, without any external dependencies. Although it is mainly designed to take advantage of massive GPU parallelism, Espresso also provides an equivalent CPU implementation for CNNs. Espresso provides special convolutional and dense layers for BCNNs, leveraging bit-packing and bit-wise computations for efficient execution. These techniques provide a speed-up of matrix-multiplication routines, and at the same time, reduce memory usage when storing parameters and activations. We experimentally show that Espresso is significantly faster than existing implementations of optimized binary neural networks (~ 2 orders of magnitude). Espresso is released under the Apache 2.0 license and is available at http://github.com/organization/project.
****

Towards Deep Learning Models Resistant to Adversarial Attacks
Aleksander Madry,Aleksandar Makelov,Ludwig Schmidt,Dimitris Tsipras,Adrian Vladu
Recent work has demonstrated that neural networks are vulnerable to adversarial examples, i.e., inputs that are almost indistinguishable from natural data and y et classified incorrectly by the network. To address this problem, we study the

adversarial robustness of neural networks through the lens of robust optimizatio
n. This approach provides us with a broad and unifying view on much prior work o
n this topic. Its principled nature also enables us to identify methods for both
 training and attacking neural networks that are reliable and, in a certain sens
e, universal. In particular, they specify a concrete security guarantee that wou
ld protect against a well-defined class of adversaries. These methods let us tra
in networks with significantly improved resistance to a wide range of adversaria
l attacks. They also suggest robustness against a first-order adversary as a nat
ural security guarantee. We believe that robustness against such well-defined cl
asses of adversaries is an important stepping stone towards fully resistant deep
 learning models.
****
Learning Deep Generative Models of Graphs
Yujia Li,Oriol Vinyals,Chris Dyer,Razvan Pascanu,Peter Battaglia
Graphs are fundamental data structures required to model many important real-wor
ld data, from knowledge graphs, physical and social interactions to molecules an
d proteins. In this paper, we study the problem of learning generative models of
 graphs from a dataset of graphs of interest. After learning, these models can b
e used to generate samples with similar properties as the ones in the dataset.
Such models can be useful in a lot of applications, e.g. drug discovery and know
ledge graph construction. The task of learning generative models of graphs, howe
ver, has its unique challenges. In particular, how to handle symmetries in graph
s and ordering of its elements during the generation process are important issue
s. We propose a generic graph neural net based model that is capable of generati
ng any arbitrary graph. We study its performance on a few graph generation task
s compared to baselines that exploit domain knowledge. We discuss potential iss
ues and open problems for such generative models going forward.
****
Covariant Compositional Networks For Learning Graphs
Risi Kondor,Truong Son Hy,Horace Pan,Brandon M. Anderson,Shubhendu Trivedi
Most existing neural networks for learning graphs deal with the issue of permuta
tion invariance by conceiving of the network as a message passing scheme, where
each node sums the feature vectors coming from its neighbors. We argue that this
 imposes a limitation on their representation power, and instead propose a new g
eneral architecture for representing objects consisting of a hierarchy of parts,
 which we call Covariant Compositional Networks (CCNs). Here covariance means th
at the activation of each neuron must transform in a specific way under permutat
ions, similarly to steerability in CNNs. We achieve covariance by making each ac
tivation transform according to a tensor representation of the permutation group
, and derive the corresponding tensor aggregation rules that each neuron must im
plement. Experiments show that CCNs can outperform competing methods on some sta
ndard graph learning benchmarks.
****
Neural Task Graph Execution
Sungryull Sohn,Junhyuk Oh,Honglak Lee
In order to develop a scalable multi-task reinforcement learning (RL) agent that
 is able to execute many complex tasks, this paper introduces a new RL problem w
here the agent is required to execute a given task graph which describes a set o
f subtasks and dependencies among them. Unlike existing approaches which explici
tly describe what the agent should do, our problem only describes properties of
subtasks and relationships between them, which requires the agent to perform a c
omplex reasoning to find the optimal subtask to execute. To solve this problem,
we propose a neural task graph solver (NTS) which encodes the task graph using a
 recursive neural network. To overcome the difficulty of training, we propose a
novel non-parametric gradient-based policy that performs back-propagation over a
 differentiable form of the task graph to compute the influence of each subtask
on the other subtasks. Our NTS is pre-trained to approximate the proposed gradie
nt-based policy and fine-tuned through actor-critic method. The experimental res
ults on a 2D visual domain show that our method to pre-train from the gradient-b
ased policy significantly improves the performance of NTS. We also demonstrate t

hat our agent can perform a complex reasoning to find the optimal way of executi
ng the task graph and generalize well to unseen task graphs. In addition, we com
pare our agent with a Monte-Carlo Tree Search (MCTS) method showing that our met
hod is much more efficient than MCTS, and the performance of our agent can be fu
rther improved by combining with MCTS. The demo video is available at https://yo
utu.be/e_ZXVS5VutM.
****

Learning to Teach

Yang Fan,Fei Tian,Tao Qin,Xiang-Yang Li,Tie-Yan Liu

Teaching plays a very important role in our society, by spreading human knowledg
e and educating our next generations. A good teacher will select appropriate tea
ching materials, impact suitable methodologies, and set up targeted examinations
, according to the learning behaviors of the students. In the field of artificia
l intelligence, however, one has not fully explored the role of teaching, and pa
ys most attention to machine \emph{learning}. In this paper, we argue that equal
 attention, if not more, should be paid to teaching, and furthermore, an optimiz
ation framework (instead of heuristics) should be used to obtain good teaching s
trategies. We call this approach ``learning to teach''. In the approach, two int
elligent agents interact with each other: a student model (which corresponds to
the learner in traditional machine learning algorithms), and a teacher model (wh
ich determines the appropriate data, loss function, and hypothesis space to faci
litate the training of the student model). The teacher model leverages the feedb
ack from the student model to optimize its own teaching strategies by means of r
einforcement learning, so as to achieve teacher-student co-evolution. To demonst
rate the practical value of our proposed approach, we take the training of deep
neural networks (DNN) as an example, and show that by using the learning to teac
h techniques, we are able to use much less training data and fewer iterations to
 achieve almost the same accuracy for different kinds of DNN models (e.g., multi
-layer perceptron, convolutional neural networks and recurrent neural networks)
under various machine learning tasks (e.g., image classification and text unders
tanding).
****

Achieving morphological agreement with Concorde

Daniil Polykovskiy,Dmitry Soloviev

Neural conversational models are widely used in applications like personal assis
tants and chat bots. These models seem to give better performance when operating
 on word level. However, for fusion languages like French, Russian and Polish vo
cabulary size sometimes become infeasible since most of the words have lots of w
ord forms. We propose a neural network architecture for transforming normalized
text into a grammatically correct one. Our model efficiently employs corresponde
nce between normalized and target words and significantly outperforms character-
level models while being 2x faster in training and 20\% faster at evaluation. We
 also propose a new pipeline for building conversational models: first generate
a normalized answer and then transform it into a grammatically correct one using
 our network. The proposed pipeline gives better performance than character-leve
l conversational models according to assessor testing.
****

Transfer Learning to Learn with Multitask Neural Model Search

Catherine Wong,Andrea Gesmundo

Deep learning models require extensive architecture design exploration and hyper
parameter optimization to perform well on a given task. The exploration of the m
odel design space is often made by a human expert, and optimized using a combina
tion of grid search and search heuristics over a large space of possible choices
. Neural Architecture Search (NAS) is a Reinforcement Learning approach that has
 been proposed to automate architecture design. NAS has been successfully applie
d to generate Neural Networks that rival the best human-designed architectures.
However, NAS requires sampling, constructing, and training hundreds to thousands
 of models to achieve well-performing architectures. This procedure needs to be
executed from scratch for each new task. The application of NAS to a wide set of
 tasks currently lacks a way to transfer generalizable knowledge across tasks.

In this paper, we present the Multitask Neural Model Search (MNMS) controller. Our goal is to learn a generalizable framework that can condition model construction on successful model searches for previously seen tasks, thus significantly speeding up the search for new tasks. We demonstrate that MNMS can conduct an automated architecture search for multiple tasks simultaneously while still learning well-performing, specialized models for each task. We then show that pre-trained MNMS controllers can transfer learning to new tasks. By leveraging knowledge from previous searches, we find that pre-trained MNMS models start from a better location in the search space and reduce search time on unseen tasks, while still discovering models that outperform published human-designed models.
****

Learning from Between-class Examples for Deep Sound Recognition
Yuji Tokozume,Yoshitaka Ushiku,Tatsuya Harada
Deep learning methods have achieved high performance in sound recognition tasks. Deciding how to feed the training data is important for further performance improvement. We propose a novel learning method for deep sound recognition: Between-Class learning (BC learning). Our strategy is to learn a discriminative feature space by recognizing the between-class sounds as between-class sounds. We generate between-class sounds by mixing two sounds belonging to different classes with a random ratio. We then input the mixed sound to the model and train the model to output the mixing ratio. The advantages of BC learning are not limited only to the increase in variation of the training data; BC learning leads to an enlargement of Fisher’s criterion in the feature space and a regularization of the positional relationship among the feature distributions of the classes. The experimental results show that BC learning improves the performance on various sound recognition networks, datasets, and data augmentation schemes, in which BC learning proves to be always beneficial. Furthermore, we construct a new deep sound recognition network (EnvNet-v2) and train it with BC learning. As a result, we achieved a performance surpasses the human level.
****

PrivyNet: A Flexible Framework for Privacy-Preserving Deep Neural Network Training
Meng Li,Liangzhen Lai,Naveen Suda,Vikas Chandra,David Z. Pan
Massive data exist among user local platforms that usually cannot support deep neural network (DNN) training due to computation and storage resource constraints. Cloud-based training schemes provide beneficial services but suffer from potential privacy risks due to excessive user data collection. To enable cloud-based DNN training while protecting the data privacy simultaneously, we propose to leverage the intermediate representations of the data, which is achieved by splitting the DNNs and deploying them separately onto local platforms and the cloud. The local neural network (NN) is used to generate the feature representations. To avoid local training and protect data privacy, the local NN is derived from pre-trained NNs. The cloud NN is then trained based on the extracted intermediate representations for the target learning task. We validate the idea of DNN splitting by characterizing the dependency of privacy loss and classification accuracy on the local NN topology for a convolutional NN (CNN) based image classification task. Based on the characterization, we further propose PrivyNet to determine the local NN topology, which optimizes the accuracy of the target learning task under the constraints on privacy loss, local computation, and storage. The efficiency and effectiveness of PrivyNet are demonstrated with CIFAR-10 dataset.
****

Multiscale Hidden Markov Models For Covariance Prediction
João Sedoc,Jordan Rodu,Dean Foster,Lyle Ungar
This paper presents a novel variant of hierarchical hidden Markov models (HMMs), the multiscale hidden Markov model (MSHMM), and an associated spectral estimation and prediction scheme that is consistent, finds global optima, and is computationally efficient. Our MSHMM is a generative model of multiple HMMs evolving at different rates where the observation is a result of the additive emissions of the HMMs. While estimation is relatively straightforward, prediction for the MSHMM poses a unique challenge, which we address in this paper. Further, we show t

hat spectral estimation of the MSHMM outperforms standard methods of predicting the asset covariance of stock prices, a widely addressed problem that is multiscale, non-stationary, and requires processing huge amounts of data.
****
Hierarchical Representations for Efficient Architecture Search
Hanxiao Liu,Karen Simonyan,Oriol Vinyals,Chrisantha Fernando,Koray Kavukcuoglu
We explore efficient neural architecture search methods and show that a simple yet powerful evolutionary algorithm can discover new architectures with excellent performance. Our approach combines a novel hierarchical genetic representation scheme that imitates the modularized design pattern commonly adopted by human experts, and an expressive search space that supports complex topologies. Our algorithm efficiently discovers architectures that outperform a large number of manually designed models for image classification, obtaining top-1 error of 3.6% on CIFAR-10 and 20.3% when transferred to ImageNet, which is competitive with the best existing neural architecture search approaches. We also present results using random search, achieving 0.3% less top-1 accuracy on CIFAR-10 and 0.1% less on ImageNet whilst reducing the search time from 36 hours down to 1 hour.
****
Distribution Regression Network
Connie Kou,Hwee Kuan Lee,Teck Khim Ng
We introduce our Distribution Regression Network (DRN) which performs regression from input probability distributions to output probability distributions. Compared to existing methods, DRN learns with fewer model parameters and easily extends to multiple input and multiple output distributions. On synthetic and real-world datasets, DRN performs similarly or better than the state-of-the-art. Furthermore, DRN generalizes the conventional multilayer perceptron (MLP). In the framework of MLP, each node encodes a real number, whereas in DRN, each node encodes a probability distribution.
****
Neural Compositional Denotational Semantics for Question Answering
Nitish Gupta,Mike Lewis
Answering compositional questions requiring multi-step reasoning is challenging for current models. We introduce an end-to-end differentiable model for interpreting questions, which is inspired by formal approaches to semantics. Each span of text is represented by a denotation in a knowledge graph, together with a vector that captures ungrounded aspects of meaning. Learned composition modules recursively combine constituents, culminating in a grounding for the complete sentence which is an answer to the question. For example, to interpret ‘not green’, the model will represent ‘green’ as a set of entities, ‘not’ as a trainable ungrounded vector, and then use this vector to parametrize a composition function to perform a complement operation. For each sentence, we build a parse chart subsuming all possible parses, allowing the model to jointly learn both the composition operators and output structure by gradient descent. We show the model can learn to represent a variety of challenging semantic operators, such as quantifiers, negation, disjunctions and composed relations on a synthetic question answering task. The model also generalizes well to longer sentences than seen in its training data, in contrast to LSTM and RelNet baselines. We will release our code.
****
Learning Latent Representations in Neural Networks for Clustering through Pseudo Supervision and Graph-based Activity Regularization
Ozsel Kilinc,Ismail Uysal
In this paper, we propose a novel unsupervised clustering approach exploiting the hidden information that is indirectly introduced through a pseudo classification objective. Specifically, we randomly assign a pseudo parent-class label to each observation which is then modified by applying the domain specific transformation associated with the assigned label. Generated pseudo observation-label pairs are subsequently used to train a neural network with Auto-clustering Output Layer (ACOL) that introduces multiple softmax nodes for each pseudo parent-class. Due to the unsupervised objective based on Graph-based Activity Regularization (GAR) terms, softmax duplicates of each parent-class are specialized as the hidde

n information captured through the help of domain specific transformations is propagated during training. Ultimately we obtain a k-means friendly latent representation. Furthermore, we demonstrate how the chosen transformation type impacts performance and helps propagate the latent information that is useful in revealing unknown clusters. Our results show state-of-the-art performance for unsupervised clustering tasks on MNIST, SVHN and USPS datasets, with the highest accuracies reported to date in the literature.
****

Deep Continuous Clustering

Sohil Atul Shah,Vladlen Koltun

Clustering high-dimensional datasets is hard because interpoint distances become less informative in high-dimensional spaces. We present a clustering algorithm that performs nonlinear dimensionality reduction and clustering jointly. The data is embedded into a lower-dimensional space by a deep autoencoder. The autoencoder is optimized as part of the clustering process. The resulting network produces clustered data. The presented approach does not rely on prior knowledge of the number of ground-truth clusters. Joint nonlinear dimensionality reduction and clustering are formulated as optimization of a global continuous objective. We thus avoid discrete reconfigurations of the objective that characterize prior clustering algorithms. Experiments on datasets from multiple domains demonstrate that the presented algorithm outperforms state-of-the-art clustering schemes, including recent methods that use deep networks.
****

Ensemble Methods as a Defense to Adversarial Perturbations Against Deep Neural Networks

Thilo Strauss,Markus Hanselmann,Andrej Junginger,Holger Ulmer

Deep learning has become the state of the art approach in many machine learning problems such as classification. It has recently been shown that deep learning is highly vulnerable to adversarial perturbations. Taking the camera systems of self-driving cars as an example, small adversarial perturbations can cause the system to make errors in important tasks, such as classifying traffic signs or detecting pedestrians. Hence, in order to use deep learning without safety concerns a proper defense strategy is required. We propose to use ensemble methods as a defense strategy against adversarial perturbations. We find that an attack leading one model to misclassify does not imply the same for other networks performing the same task. This makes ensemble methods an attractive defense strategy against adversarial attacks. We empirically show for the MNIST and the CIFAR-10 data sets that ensemble methods not only improve the accuracy of neural networks on test data but also increase their robustness against adversarial perturbations.
****

Divide-and-Conquer Reinforcement Learning

Dibya Ghosh,Avi Singh,Aravind Rajeswaran,Vikash Kumar,Sergey Levine

Standard model-free deep reinforcement learning (RL) algorithms sample a new initial state for each trial, allowing them to optimize policies that can perform well even in highly stochastic environments. However, problems that exhibit considerable initial state variation typically produce high-variance gradient estimates for model-free RL, making direct policy or value function optimization challenging. In this paper, we develop a novel algorithm that instead partitions the initial state space into "slices", and optimizes an ensemble of policies, each on a different slice. The ensemble is gradually unified into a single policy that can succeed on the whole state space. This approach, which we term divide-and-conquer RL, is able to solve complex tasks where conventional deep RL methods are ineffective. Our results show that divide-and-conquer RL greatly outperforms conventional policy gradient methods on challenging grasping, manipulation, and locomotion tasks, and exceeds the performance of a variety of prior methods. Videos of policies learned by our algorithm can be viewed at https://sites.google.com/view/dnc-rl/

****

Predicting Multiple Actions for Stochastic Continuous Control

Sanjeev Kumar,Christian Rupprecht,Federico Tombari,Gregory D. Hager
We introduce a new approach to estimate continuous actions using actor-critic al
gorithms for reinforcement learning problems. Policy gradient methods usually pr
edict one continuous action estimate or parameters of a presumed distribution (m
ost commonly Gaussian) for any given state which might not be optimal as it may
not capture the complete description of the target distribution. Our approach in
stead predicts M actions with the policy network (actor) and then uniformly samp
le one action during training as well as testing at each state. This allows the
agent to learn a simple stochastic policy that has an easy to compute expected r
eturn. In all experiments, this facilitates better exploration of the state spac
e during training and converges to a better policy.
****
Log-DenseNet: How to Sparsify a DenseNet
Hanzhang Hu,Debadeepta Dey,Allie Del Giorno,Martial Hebert,J. Andrew Bagnell
Skip connections are increasingly utilized by deep neural networks to improve ac
curacy and cost-efficiency. In particular, the recent DenseNet is efficient in c
omputation and parameters, and achieves state-of-the-art predictions by directly
 connecting each feature layer to all previous ones. However, DenseNet's extreme
 connectivity pattern may hinder its scalability to high depths, and in applicat
ions like fully convolutional networks, full DenseNet connections are prohibitiv
ely expensive.
This work first experimentally shows that one key advantage of skip connections
is to have short distances among feature layers during backpropagation. Specific
ally, using a fixed number of skip connections, the connection patterns with sho
rter backpropagation distance among layers have more accurate predictions. Follo
wing this insight, we propose a connection template, Log-DenseNet, which, in com
parison to DenseNet, only slightly increases the backpropagation distances amon
g layers from 1 to ($1 + \log_2 L$), but uses only $L\log_2 L$ total connection
s instead of $O(L^2)$. Hence, \logdenses are easier to scale than DenseNets, and
 no longer require careful GPU memory management. We demonstrate the effectivene
ss of our design principle by showing better performance than DenseNets on tabul
a rasa semantic segmentation, and competitive results on visual recognition.
****
N2N learning: Network to Network Compression via Policy Gradient Reinforcement L
earning
Anubhav Ashok,Nicholas Rhinehart,Fares Beainy,Kris M. Kitani
While bigger and deeper neural network architectures continue to advance the sta
te-of-the-art for many computer vision tasks, real-world adoption of these netwo
rks is impeded by hardware and speed constraints. Conventional model compression
 methods attempt to address this problem by modifying the architecture manually
or using pre-defined heuristics. Since the space of all reduced architectures is
 very large, modifying the architecture of a deep neural network in this way is
a difficult task. In this paper, we tackle this issue by introducing a principle
d method for learning reduced network architectures in a data-driven way using r
einforcement learning. Our approach takes a larger 'teacher' network as input an
d outputs a compressed 'student' network derived from the 'teacher' network. In
the first stage of our method, a recurrent policy network aggressively removes l
ayers from the large 'teacher' model. In the second stage, another recurrent po
licy network carefully reduces the size of each remaining layer. The resulting n
etwork is then evaluated to obtain a reward -- a score based on the accuracy and
 compression of the network. Our approach uses this reward signal with policy gr
adients to train the policies to find a locally optimal student network. Our exp
eriments show that we can achieve compression rates of more than 10x for models
such as ResNet-34 while maintaining similar performance to the input 'teacher' n
etwork. We also present a valuable transfer learning result which shows that pol
icies which are pre-trained on smaller 'teacher' networks can be used to rapidly
 speed up training on larger 'teacher' networks.
****
Gaussian Process Behaviour in Wide Deep Neural Networks
Alexander G. de G. Matthews,Jiri Hron,Mark Rowland,Richard E. Turner,Zoubin Ghah

ramani
Whilst deep neural networks have shown great empirical success, there is still much work to be done to understand their theoretical properties. In this paper, we study the relationship between Gaussian processes with a recursive kernel definition and random wide fully connected feedforward networks with more than one hidden layer. We exhibit limiting procedures under which finite deep networks will converge in distribution to the corresponding Gaussian process. To evaluate convergence rates empirically, we use maximum mean discrepancy. We then exhibit situations where existing Bayesian deep networks are close to Gaussian processes in terms of the key quantities of interest. Any Gaussian process has a flat representation. Since this behaviour may be undesirable in certain situations we discuss ways in which it might be prevented.
****
Alternating Multi-bit Quantization for Recurrent Neural Networks
Chen Xu,Jianqiang Yao,Zhouchen Lin,Wenwu Ou,Yuanbin Cao,Zhirong Wang,Hongbin Zha
Recurrent neural networks have achieved excellent performance in many applications. However, on portable devices with limited resources, the models are often too large to deploy. For applications on the server with large scale concurrent requests, the latency during inference can also be very critical for costly computing resources. In this work, we address these problems by quantizing the network, both weights and activations, into multiple binary codes {-1,+1}. We formulate the quantization as an optimization problem. Under the key observation that once the quantization coefficients are fixed the binary codes can be derived efficiently by binary search tree, alternating minimization is then applied. We test the quantization for two well-known RNNs, i.e., long short term memory (LSTM) and gated recurrent unit (GRU), on the language models. Compared with the full-precision counter part, by 2-bit quantization we can achieve ~16x memory saving and ~6x real inference acceleration on CPUs, with only a reasonable loss in the accuracy. By 3-bit quantization, we can achieve almost no loss in the accuracy or even surpass the original model, with ~10.5x memory saving and ~3x real inference acceleration. Both results beat the exiting quantization works with large margins. We extend our alternating quantization to image classification tasks. In both RNNs and feedforward neural networks, the method also achieves excellent performance.
****
Incremental Learning through Deep Adaptation
Amir Rosenfeld,John K. Tsotsos
Given an existing trained neural network, it is often desirable to learn new capabilities without hindering performance of those already learned. Existing approaches either learn sub-optimal solutions, require joint training, or incur a substantial increment in the number of parameters for each added task, typically as many as the original network. We propose a method called Deep Adaptation Networks (DAN) that constrains newly learned filters to be linear combinations of existing ones. DANs preserve performance on the original task, require a fraction (typically 13%) of the number of parameters compared to standard fine-tuning procedures and converge in less cycles of training to a comparable or better level of performance. When coupled with standard network quantization techniques, we further reduce the parameter cost to around 3% of the original with negligible or no loss in accuracy. The learned architecture can be controlled to switch between various learned representations, enabling a single network to solve a task from multiple different domains. We conduct extensive experiments showing the effectiveness of our method on a range of image classification tasks and explore different aspects of its behavior.
****
Divide and Conquer Networks
Alex Nowak,David Folqué,Joan Bruna
We consider the learning of algorithmic tasks by mere observation of input-output
pairs. Rather than studying this as a black-box discrete regression problem with no assumption whatsoever on the input-output mapping, we concentrate on tasks

that are amenable to the principle of divide and conquer, and study what are its
implications in terms of learning.
This principle creates a powerful inductive bias that we leverage with neural
architectures that are defined recursively and dynamically, by learning two scal
e-
invariant atomic operations: how to split a given input into smaller sets, and h
ow
to merge two partially solved tasks into a larger partial solution. Our model ca
n be
trained in weakly supervised environments, namely by just observing input-output
pairs, and in even weaker environments, using a non-differentiable reward signal
.
Moreover, thanks to the dynamic aspect of our architecture, we can incorporate
the computational complexity as a regularization term that can be optimized by
backpropagation. We demonstrate the flexibility and efficiency of the Divide-
and-Conquer Network on several combinatorial and geometric tasks: convex hull,
clustering, knapsack and euclidean TSP. Thanks to the dynamic programming
nature of our model, we show significant improvements in terms of generalization
error and computational complexity.
****
Lung Tumor Location and Identification with AlexNet and a Custom CNN
Allison M Rossetto,Wenjin Zhou
Lung cancer is the leading cause of cancer deaths in the world and early detecti
on is a crucial part of increasing patient survival. Deep learning techniques pr
ovide us with a method of automated analysis of patient scans. In this work, we
compare AlexNet, a multi-layered and highly ■exible architecture, with a custom
CNN to determine if lung nodules with patient scans are benign or cancerous. We
have found our CNN architecture to be highly accurate (99.79%) and fast while ma
intaining low False Positive and False Negative rates (< 0.01% and 0.15% respect
ively). This is important as high false positive rates are a serious issue with
lung cancer diagnosis. We have found that AlexNet is not well suited to the prob
lem of nodule identi■cation, though it is a good baseline comparison because of
its ■exibility.
****
Large Scale Multi-Domain Multi-Task Learning with MultiModel
Lukasz Kaiser,Aidan N. Gomez,Noam Shazeer,Ashish Vaswani,Niki Parmar,Llion Jones
,Jakob Uszkoreit
Deep learning yields great results across many fields,
from speech recognition, image classification, to translation.
But for each problem, getting a deep model to work well involves
research into the architecture and a long period of tuning.

We present a single model that yields good results on a number
of problems spanning multiple domains. In particular, this single model
is trained concurrently on ImageNet, multiple translation tasks,
image captioning (COCO dataset), a speech recognition corpus,
and an English parsing task.

Our model architecture incorporates building blocks from multiple
domains. It contains convolutional layers, an attention mechanism,
and sparsely-gated layers.

Each of these computational blocks is crucial for a subset of
the tasks we train on. Interestingly, even if a block is not
crucial for a task, we observe that adding it never hurts performance
and in most cases improves it on all tasks.

We also show that tasks with less data benefit largely from joint
training with other tasks, while performance on large tasks degrades
only slightly if at all.

****
Backpropagation through the Void: Optimizing control variates for black-box grad
ient estimation
Will Grathwohl,Dami Choi,Yuhuai Wu,Geoff Roeder,David Duvenaud
Gradient-based optimization is the foundation of deep learning and reinforcement
 learning.
Even when the mechanism being optimized is unknown or not differentiable, optimi
zation using high-variance or biased gradient estimates is still often the best
strategy. We introduce a general framework for learning low-variance, unbiased g
radient estimators for black-box functions of random variables, based on gradien
ts of a learned function.
These estimators can be jointly trained with model parameters or policies, and a
re applicable in both discrete and continuous settings. We give unbiased, adapti
ve analogs of state-of-the-art reinforcement learning methods such as advantage
actor-critic. We also demonstrate this framework for training discrete latent-va
riable models.
****
Understanding Grounded Language Learning Agents
Felix Hill,Karl Moritz Hermann,Phil Blunsom,Stephen Clark
Neural network-based systems can now learn to locate the referents of words and
phrases in images, answer questions about visual scenes, and even execute symbol
ic instructions as first-person actors in partially-observable worlds. To achiev
e this so-called grounded language learning, models must overcome certain well-s
tudied learning challenges that are also fundamental to infants learning their f
irst words. While it is notable that models with no meaningful prior knowledge o
vercome these learning obstacles, AI researchers and practitioners currently lac
k a clear understanding of exactly how they do so. Here we address this question
 as a way of achieving a clearer general understanding of grounded language lear
ning, both to inform future research and to improve confidence in model predicti
ons. For maximum control and generality, we focus on a simple neural network-bas
ed language learning agent trained via policy-gradient methods to interpret synt
hetic linguistic instructions in a simulated 3D world. We apply experimental par
adigms from developmental psychology to this agent, exploring the conditions und
er which established human biases and learning effects emerge. We further propos
e a novel way to visualise and analyse semantic representation in grounded langu
age learning agents that yields a plausible computational account of the observe
d effects.
****
Massively Parallel Hyperparameter Tuning
Lisha Li,Kevin Jamieson,Afshin Rostamizadeh,Katya Gonina,Moritz Hardt,Benjamin R
echt,Ameet Talwalkar
Modern machine learning models are characterized by large hyperparameter search
spaces and prohibitively expensive training costs. For such models, we cannot a
fford to train candidate models sequentially and wait months before finding a su
itable hyperparameter configuration. Hence, we introduce the large-scale regime
for parallel hyperparameter tuning, where we need to evaluate orders of magnitud
e more configurations than available parallel workers in a small multiple of the
 wall-clock time needed to train a single model. We propose a novel hyperparame
ter tuning algorithm for this setting that exploits both parallelism and aggress
ive early-stopping techniques, building on the insights of the Hyperband algorit
hm. Finally, we conduct a thorough empirical study of our algorithm on several
benchmarks, including large-scale experiments with up to 500 workers. Our resul
ts show that our proposed algorithm finds good hyperparameter settings nearly an
 order of magnitude faster than random search.
****
Progressive Growing of GANs for Improved Quality, Stability, and Variation
Tero Karras,Timo Aila,Samuli Laine,Jaakko Lehtinen
We describe a new training methodology for generative adversarial networks. The
key idea is to grow both the generator and discriminator progressively: starting
 from a low resolution, we add new layers that model increasingly fine details a

s training progresses. This both speeds the training up and greatly stabilizes i
t, allowing us to produce images of unprecedented quality, e.g., CelebA images a
t 1024^2. We also propose a simple way to increase the variation in generated im
ages, and achieve a record inception score of 8.80 in unsupervised CIFAR10. Addi
tionally, we describe several implementation details that are important for disc
ouraging unhealthy competition between the generator and discriminator. Finally,
 we suggest a new metric for evaluating GAN results, both in terms of image qual
ity and variation. As an additional contribution, we construct a higher-quality
version of the CelebA dataset.
****
Synthetic and Natural Noise Both Break Neural Machine Translation
Yonatan Belinkov,Yonatan Bisk
Character-based neural machine translation (NMT) models alleviate out-of-vocabul
ary issues, learn morphology, and move us closer to completely end-to-end transl
ation systems. Unfortunately, they are also very brittle and easily falter when
 presented with noisy data. In this paper, we confront NMT models with synthetic
 and natural sources of noise. We find that state-of-the-art models fail to tran
slate even moderately noisy texts that humans have no trouble comprehending. We
explore two approaches to increase model robustness: structure-invariant word re
presentations and robust training on noisy texts. We find that a model based on
a character convolutional neural network is able to simultaneously learn represe
ntations robust to multiple kinds of noise.
****
Representing dynamically: An active process for describing sequential data
Juan Sebastian Olier,Emilia Barakova,Matthias Rauterberg,Carlo Regazzoni
We propose an unsupervised method for building dynamic representations of sequen
tial data, particularly of observed interactions. The method simultaneously acqu
ires representations of input data and its dynamics. It is based on a hierarchic
al generative model composed of two levels. In the first level, a model learns r
epresentations to generate observed data. In the second level, representational
states encode the dynamics of the lower one. The model is designed as a Bayesian
 network with switching variables represented in the higher level, and which gen
erates transition models. The method actively explores the latent space guided b
y its knowledge and the uncertainty about it. That is achieved by updating the l
atent variables from prediction error signals backpropagated to the latent space
. So, no encoder or inference models are used since the generators also serve as
 their inverse transformations.
The method is evaluated in two scenarios, with static images and with videos. Th
e results show that the adaptation over time leads to better performance than wi
th similar architectures without temporal dependencies, e.g., variational autoen
coders. With videos, it is shown that the system extracts the dynamics of the da
ta in states that highly correlate with the ground truth of the actions observed
.
****
Universality, Robustness, and Detectability of Adversarial Perturbations under A
dversarial Training
Jan Hendrik Metzen
Classifiers such as deep neural networks have been shown to be vulnerable agains
t adversarial perturbations on problems with high-dimensional input space. While
 adversarial training improves the robustness of classifiers against such advers
arial perturbations, it leaves classifiers sensitive to them on a non-negligible
 fraction of the inputs. We argue that there are two different kinds of adversar
ial perturbations: shared perturbations which fool a classifier on many inputs a
nd singular perturbations which only fool the classifier on a small fraction of
the data. We find that adversarial training increases the robustness of classifi
ers against shared perturbations. Moreover, it is particularly effective in remo
ving universal perturbations, which can be seen as an extreme form of shared per
turbations. Unfortunately, adversarial training does not consistently increase t
he robustness against singular perturbations on unseen inputs. However, we find
that adversarial training decreases robustness of the remaining perturbations ag

ainst image transformations such as changes to contrast and brightness or Gauss
ian blurring. It thus makes successful attacks on the classifier in the physical
 world less likely. Finally, we show that even singular perturbations can be eas
ily detected and must thus exhibit generalizable patterns even though the pertur
bations are specific for certain inputs.
****

Generative Entity Networks: Disentangling Entitites and Attributes in Visual Sce
nes using Partial Natural Language Descriptions
Charlie Nash,Sebastian Nowozin,Nate Kushman
Generative image models have made significant progress in the last few years, an
d are now able to generate low-resolution images which sometimes look realistic.
 However the state-of-the-art models utilize fully entangled latent representati
ons where small changes to a single neuron can effect every output pixel in rela
tively arbitrary ways, and different neurons have possibly arbitrary relationshi
ps with each other. This limits the ability of such models to generalize to new
combinations or orientations of objects as well as their ability to connect with
 more structured representations such as natural language, without explicit stro
ng supervision. In this work explore the synergistic effect of using partial nat
ural language scene descriptions to help disentangle the latent entities visible
 an image. We present a novel neural network architecture called Generative Enti
ty Networks, which jointly generates both the natural language descriptions and
the images from a set of latent entities. Our model is based on the variational
autoencoder framework and makes use of visual attention to identify and characte
rise the visual attributes of each entity. Using the Shapeworld dataset, we show
 that our representation both enables a better generative model of images, leadi
ng to higher quality image samples, as well as creating more semantically useful
 representations that improve performance over purely dicriminative models on a
simple natural language yes/no question answering task.
****

Zero-shot Cross Language Text Classification
Dan Svenstrup,Jonas Meinertz Hansen,Ole Winther
Labeled text classification datasets are typically only available in a few selec
t languages. In order to train a model for e.g news categorization in a language
 L_t without a suitable text classification dataset there are two options. The
 first option is to create a new labeled dataset by hand, and the second option
is to transfer label information from an existing labeled dataset in a source la
nguage L_s to the target language L_t. In this paper we propose a method for
 sharing label information across languages by means of a language independent t
ext encoder. The encoder will give almost identical representations to multiling
ual versions of the same text. This means that labeled data in one language can
be used to train a classifier that works for the rest of the languages. The enco
der is trained independently of any concrete classification task and can therefo
re subsequently be used for any classification task. We show that it is possibl
e to obtain good performance even in the case where only a comparable corpus of
texts is available.
****

UCB EXPLORATION VIA Q-ENSEMBLES
Richard Y. Chen,Szymon Sidor,Pieter Abbeel,John Schulman
We show how an ensemble of Q^*-functions can be leveraged for more effective e
xploration in deep reinforcement learning. We build on well established algorith
ms from the bandit setting, and adapt them to the Q-learning setting. We propo
se an exploration strategy based on upper-confidence bounds (UCB). Our experimen
ts show significant gains on the Atari benchmark.
****

BLOCK-NORMALIZED GRADIENT METHOD: AN EMPIRICAL STUDY FOR TRAINING DEEP NEURAL NE
TWORK
Adams Wei Yu,Lei Huang,Qihang Lin,Ruslan Salakhutdinov,Jaime Carbonell
In this paper, we propose a generic and simple strategy for utilizing stochastic
 gradient information in optimization. The technique essentially contains two co
nsecutive steps in each iteration: 1) computing and normalizing each block (laye

r) of the mini-batch stochastic gradient; 2) selecting appropriate step size to update the decision variable (parameter) towards the negative of the block-normalized gradient. We conduct extensive empirical studies on various non-convex neural network optimization problems, including multilayer perceptron, convolution neural networks and recurrent neural networks. The results indicate the block-normalized gradient can help accelerate the training of neural networks. In particular,
we observe that the normalized gradient methods having constant step size with occasionally decay, such as SGD with momentum, have better performance in the deep convolution neural networks, while those with adaptive step sizes, such as Adam, perform better in recurrent neural networks. Besides, we also observe this line of methods can lead to solutions with better generalization properties, which is confirmed by the performance improvement over strong baselines.
****
Neural Networks for irregularly observed continuous-time Stochastic Processes
Francois W. Belletti,Alexander Ku,Joseph E. Gonzalez
Designing neural networks for continuous-time stochastic processes is challenging, especially when observations are made irregularly. In this article, we analyze neural networks from a frame theoretic perspective to identify the sufficient conditions that enable smoothly recoverable representations of signals in L^2(R). Moreover, we show that, under certain assumptions, these properties hold even when signals are irregularly observed. As we converge to the family of (convolutional) neural networks that satisfy these conditions, we show that we can optimize our convolution filters while constraining them so that they effectively compute a Discrete Wavelet Transform. Such a neural network can efficiently divide the time-axis of a signal into orthogonal sub-spaces of different temporal scale and localization. We evaluate the resulting neural network on an assortment of synthetic and real-world tasks: parsimonious auto-encoding, video classification, and financial forecasting.
****
Stable Distribution Alignment Using the Dual of the Adversarial Distance
Ben Usman,Kate Saenko,Brian Kulis
Methods that align distributions by minimizing an adversarial distance between them have recently achieved impressive results. However, these approaches are difficult to optimize with gradient descent and they often do not converge well without careful hyperparameter tuning and proper initialization. We investigate whether turning the adversarial min-max problem into an optimization problem by replacing the maximization part with its dual improves the quality of the resulting alignment and explore its connections to Maximum Mean Discrepancy. Our empirical results suggest that using the dual formulation for the restricted family of linear discriminators results in a more stable convergence to a desirable solution when compared with the performance of a primal min-max GAN-like objective and an MMD objective under the same restrictions. We test our hypothesis on the problem of aligning two synthetic point clouds on a plane and on a real-image domain adaptation problem on digits. In both cases, the dual formulation yields an iterative procedure that gives more stable and monotonic improvement over time.
****
Modular Continual Learning in a Unified Visual Environment
Kevin T. Feigelis,Blue Sheffer,Daniel L. K. Yamins
A core aspect of human intelligence is the ability to learn new tasks quickly and switch between them flexibly. Here, we describe a modular continual reinforcement learning paradigm inspired by these abilities. We first introduce a visual interaction environment that allows many types of tasks to be unified in a single framework. We then describe a reward map prediction scheme that learns new tasks robustly in the very large state and action spaces required by such an environment. We investigate how properties of module architecture influence efficiency of task learning, showing that a module motif incorporating specific design principles (e.g. early bottlenecks, low-order polynomial nonlinearities, and symmetry) significantly outperforms more standard neural network motifs, needing fewer training examples and fewer neurons to achieve high levels of performance. Fina

lly, we present a meta-controller architecture for task switching based on a dynamic neural voting scheme, which allows new modules to use information learned from previously-seen tasks to substantially improve their own learning efficiency.
****
Learning to diagnose from scratch by exploiting dependencies among labels
Li Yao,Eric Poblenz,Dmitry Dagunts,Ben Covington,Devon Bernard,Kevin Lyman
The field of medical diagnostics contains a wealth of challenges which closely resemble classical machine learning problems; practical constraints, however, complicate the translation of these endpoints naively into classical architectures. Many tasks in radiology, for example, are largely problems of multi-label classification wherein medical images are interpreted to indicate multiple present or suspected pathologies. Clinical settings drive the necessity for high accuracy simultaneously across a multitude of pathological outcomes and greatly limit the utility of tools which consider only a subset. This issue is exacerbated by a general scarcity of training data and maximizes the need to extract clinically relevant features from available samples -- ideally without the use of pre-trained models which may carry forward undesirable biases from tangentially related tasks. We present and evaluate a partial solution to these constraints in using LSTMs to leverage interdependencies among target labels in predicting 14 pathologic patterns from chest x-rays and establish state of the art results on the largest publicly available chest x-ray dataset from the NIH without pre-training. Furthermore, we propose and discuss alternative evaluation metrics and their relevance in clinical practice.
****
Classifier-to-Generator Attack: Estimation of Training Data Distribution from Classifier
Kosuke Kusano,Jun Sakuma
Suppose a deep classification model is trained with samples that need to be kept private for privacy or confidentiality reasons. In this setting, can an adversary obtain the private samples if the classification model is given to the adversary? We call this reverse engineering against the classification model the Classifier-to-Generator (C2G) Attack. This situation arises when the classification model is embedded into mobile devices for offline prediction (e.g., object recognition for the automatic driving car and face recognition for mobile phone authentication).
For C2G attack, we introduce a novel GAN, PreImageGAN. In PreImageGAN, the generator is designed to estimate the the sample distribution conditioned by the preimage of classification model f, $P(X|f(X)=y)$, where X is the random variable on the sample space and y is the probability vector representing the target label arbitrary specified by the adversary. In experiments, we demonstrate PreImageGAN works successfully with hand-written character recognition and face recognition. In character recognition, we show that, given a recognition model of hand-written digits, PreImageGAN allows the adversary to extract alphabet letter images without knowing that the model is built for alphabet letter images. In face recognition, we show that, when an adversary obtains a face recognition model for a set of individuals, PreImageGAN allows the adversary to extract face images of specific individuals contained in the set, even when the adversary has no knowledge of the face of the individuals.
****
Learning Sparse Structured Ensembles with SG-MCMC and Network Pruning
Yichi Zhang,Zhijian Ou
An ensemble of neural networks is known to be more robust and accurate than an individual network, however usually with linearly-increased cost in both training and testing.
In this work, we propose a two-stage method to learn Sparse Structured Ensembles (SSEs) for neural networks.
In the first stage, we run SG-MCMC with group sparse priors to draw an ensemble of samples from the posterior distribution of network parameters. In the second stage, we apply weight-pruning to each sampled network and then perform retraini

ng over the remained connections.
In this way of learning SSEs with SG-MCMC and pruning, we not only achieve high prediction accuracy since SG-MCMC enhances exploration of the model-parameter space, but also reduce memory and computation cost significantly in both training and testing of NN ensembles.
This is thoroughly evaluated in the experiments of learning SSE ensembles of both FNNs and LSTMs.
For example, in LSTM based language modeling (LM), we obtain 21\% relative reduction in LM perplexity by learning a SSE of 4 large LSTM models, which has only 30\% of model parameters and 70\% of computations in total, as compared to the baseline large LSTM LM.
To the best of our knowledge, this work represents the first methodology and empirical study of integrating SG-MCMC, group sparse prior and network pruning together for learning NN ensembles.
****
VoiceLoop: Voice Fitting and Synthesis via a Phonological Loop
Yaniv Taigman,Lior Wolf,Adam Polyak,Eliya Nachmani
We present a new neural text to speech (TTS) method that is able to transform text to speech in voices that are sampled in the wild. Unlike other systems, our solution is able to deal with unconstrained voice samples and without requiring aligned phonemes or linguistic features. The network architecture is simpler than those in the existing literature and is based on a novel shifting buffer working memory. The same buffer is used for estimating the attention, computing the output audio, and for updating the buffer itself. The input sentence is encoded using a context-free lookup table that contains one entry per character or phoneme. The speakers are similarly represented by a short vector that can also be fitted to new identities, even with only a few samples. Variability in the generated speech is achieved by priming the buffer prior to generating the audio. Experimental results on several datasets demonstrate convincing capabilities, making TTS accessible to a wider range of applications. In order to promote reproducibility, we release our source code and models.
****
Cascade Adversarial Machine Learning Regularized with a Unified Embedding
Taesik Na,Jong Hwan Ko,Saibal Mukhopadhyay
Injecting adversarial examples during training, known as adversarial training, can improve robustness against one-step attacks, but not for unknown iterative attacks. To address this challenge, we first show iteratively generated adversarial images easily transfer between networks trained with the same strategy. Inspired by this observation, we propose cascade adversarial training, which transfers the knowledge of the end results of adversarial training. We train a network from scratch by injecting iteratively generated adversarial images crafted from already defended networks in addition to one-step adversarial images from the network being trained. We also propose to utilize embedding space for both classification and low-level (pixel-level) similarity learning to ignore unknown pixel level perturbation. During training, we inject adversarial images without replacing their corresponding clean images and penalize the distance between the two embeddings (clean and adversarial). Experimental results show that cascade adversarial training together with our proposed low-level similarity learning efficiently enhances the robustness against iterative attacks, but at the expense of decreased robustness against one-step attacks. We show that combining those two techniques can also improve robustness under the worst case black box attack scenario.
****
AANN: Absolute Artificial Neural Network
Animesh Karnewar
This research paper describes a simplistic architecture named as AANN: Absolute Artificial Neural Network, which can be used to create highly interpretable representations of the input data. These representations are generated by penalizing the learning of the network in such a way that those learned representations correspond to the respective labels present in the labelled dataset used for super

vised training; thereby, simultaneously giving the network the ability to classify the input data. The network can be used in the reverse direction to generate data that closely resembles the input by feeding in representation vectors as required. This research paper also explores the use of mathematical abs (absolute valued) functions as activation functions which constitutes the core part of this neural network architecture. Finally the results obtained on the MNIST dataset by using this technique are presented and discussed in brief.
****
A Deep Predictive Coding Network for Learning Latent Representations
Shirin Dora,Cyriel Pennartz,Sander Bohte
It has been argued that the brain is a prediction machine that continuously learns how to make better predictions about the stimuli received from the external environment. For this purpose, it builds a model of the world around us and uses this model to infer the external stimulus. Predictive coding has been proposed as a mechanism through which the brain might be able to build such a model of the external environment. However, it is not clear how predictive coding can be used to build deep neural network models of the brain while complying with the architectural constraints imposed by the brain. In this paper, we describe an algorithm to build a deep generative model using predictive coding that can be used to infer latent representations about the stimuli received from external environment. Specifically, we used predictive coding to train a deep neural network on real-world images in a unsupervised learning paradigm. To understand the capacity of the network with regards to modeling the external environment, we studied the latent representations generated by the model on images of objects that are never presented to the model during training. Despite the novel features of these objects the model is able to infer the latent representations for them. Furthermore, the reconstructions of the original images obtained from these latent representations preserve the important details of these objects.
****
Block-Sparse Recurrent Neural Networks
Sharan Narang,Eric Undersander,Gregory Diamos
Recurrent Neural Networks (RNNs) are used in state-of-the-art models in domains such as speech recognition, machine translation, and language modelling. Sparsity is a technique to reduce compute and memory requirements of deep learning models. Sparse RNNs are easier to deploy on devices and high-end server processors. Even though sparse operations need less compute and memory relative to their dense counterparts, the speed-up observed by using sparse operations is less than expected on different hardware platforms. In order to address this issue, we investigate two different approaches to induce block sparsity in RNNs: pruning blocks of weights in a layer and using group lasso regularization with pruning to create blocks of weights with zeros. Using these techniques, we can create block-sparse RNNs with sparsity ranging from 80% to 90% with a small loss in accuracy. This technique allows us to reduce the model size by roughly 10x. Additionally, we can prune a larger dense network to recover this loss in accuracy while maintaining high block sparsity and reducing the overall parameter count. Our technique works with a variety of block sizes up to 32x32. Block-sparse RNNs eliminate overheads related to data storage and irregular memory accesses while increasing hardware efficiency compared to unstructured sparsity.

****
Rotational Unit of Memory
Rumen Dangovski,Li Jing,Marin Soljacic
The concepts of unitary evolution matrices and associative memory have boosted the field of Recurrent Neural Networks (RNN) to state-of-the-art performance in a variety of sequential tasks. However, RNN still has a limited capacity to manipulate long-term memory. To bypass this weakness the most successful applications of RNN use external techniques such as attention mechanisms. In this paper we propose a novel RNN model that unifies the state-of-the-art approaches: Rotational Unit of Memory (RUM). The core of RUM is its rotational operation, which is, naturally, a unitary matrix, providing architectures with the power to learn

long-term dependencies by overcoming the vanishing and exploding gradients probl em. Moreover, the rotational unit also serves as associative memory. We evalua te our model on synthetic memorization, question answering and language modeling tasks. RUM learns the Copying Memory task completely and improves the state-o f-the-art result in the Recall task. RUM's performance in the bAbI Question Ans wering task is comparable to that of models with attention mechanism. We also im prove the state-of-the-art result to 1.189 bits-per-character (BPC) loss in the Character Level Penn Treebank (PTB) task, which is to signify the applications o f RUM to real-world sequential data. The universality of our construction, at th e core of RNN, establishes RUM as a promising approach to language modeling, spe ech recognition and machine translation.
****
Towards a Testable Notion of Generalization for Generative Adversarial Networks
Robert Cornish,Hongseok Yang,Frank Wood
We consider the question of how to assess generative adversarial networks, in pa rticular with respect to whether or not they generalise beyond memorising the tr aining data. We propose a simple procedure for assessing generative adversarial network performance based on a principled consideration of what the actual goal of generalisation is. Our approach involves using a test set to estimate the Was serstein distance between the generative distribution produced by our procedure, and the underlying data distribution. We use this procedure to assess the perfo rmance of several modern generative adversarial network architectures. We find t hat this procedure is sensitive to the choice of ground metric on the underlying data space, and suggest a choice of ground metric that substantially improves p erformance. We finally suggest that attending to the ground metric used in Wass erstein generative adversarial network training may be fruitful, and outline a c oncrete pathway towards doing so.
****
Topology Adaptive Graph Convolutional Networks
Jian Du,Shanghang Zhang,Guanhang Wu,José M. F. Moura,Soummya Kar
Convolution acts as a local feature extractor in convolutional neural networks (CNNs). However, the convolution operation is not applicable when the input data is supported on an irregular graph such as with social networks, citation networ ks, or knowledge graphs. This paper proposes the topology adaptive graph convolu tional network (TAGCN), a novel graph convolutional network that generalizes CNN architectures to graph-structured data and provides a systematic way to design a set of fixed-size learnable filters to perform convolutions on graphs. The top ologies of these filters are adaptive to the topology of the graph when they sca n the graph to perform convolution, replacing the square filter for the grid-str uctured data in traditional CNNs. The outputs are the weighted sum of these filt ers' outputs, extraction of both vertex features and strength of correlation bet ween vertices. It
can be used with both directed and undirected graphs. The proposed TAGCN not onl y inherits the properties of convolutions in CNN for grid-structured data, but i t is also consistent with convolution as defined in graph signal processing. Fur ther, as no approximation to the convolution is needed, TAGCN exhibits better pe rformance than existing graph-convolution-approximation methods on a number of data sets. As only the polynomials of degree two of the adjacency matrix are used, TAGCN is also computationally simpler than other recent methods.
****
TRAINING GENERATIVE ADVERSARIAL NETWORKS VIA PRIMAL-DUAL SUBGRADIENT METHODS: A LAGRANGIAN PERSPECTIVE ON GAN
Xu Chen,Jiang Wang,Hao Ge
We relate the minimax game of generative adversarial networks (GANs) to finding the saddle points of the Lagrangian function for a convex optimization problem, where the discriminator outputs and the distribution of generator outputs play t he roles of primal variables and dual variables, respectively. This formulation shows the connection between the standard GAN training process and the primal-du al subgradient methods for convex optimization. The inherent connection does not only provide a theoretical convergence proof for training GANs in the function

space, but also inspires a novel objective function for training. The modified o
bjective function forces the distribution of generator outputs to be updated alo
ng the direction according to the primal-dual subgradient methods. A toy example
 shows that the proposed method is able to resolve mode collapse, which in this
case cannot be avoided by the standard GAN or Wasserstein GAN. Experiments on bo
th Gaussian mixture synthetic data and real-world image datasets demonstrate the
 performance of the proposed method on generating diverse samples.
****

Normalized Direction-preserving Adam
Zijun Zhang,Lin Ma,Zongpeng Li,Chuan Wu
Optimization algorithms for training deep models not only affects the convergenc
e rate and stability of the training process, but are also highly related to the
 generalization performance of trained models. While adaptive algorithms, such a
s Adam and RMSprop, have shown better optimization performance than stochastic g
radient descent (SGD) in many scenarios, they often lead to worse generalization
 performance than SGD, when used for training deep neural networks (DNNs). In th
is work, we identify two problems regarding the direction and step size for upda
ting the weight vectors of hidden units, which may degrade the generalization pe
rformance of Adam. As a solution, we propose the normalized direction-preserving
 Adam (ND-Adam) algorithm, which controls the update direction and step size mor
e precisely, and thus bridges the generalization gap between Adam and SGD. Follo
wing a similar rationale, we further improve the generalization performance in c
lassification tasks by regularizing the softmax logits. By bridging the gap betw
een SGD and Adam, we also shed some light on why certain optimization algorithms
 generalize better than others.
****

Gating out sensory noise in a spike-based Long Short-Term Memory network
Davide Zambrano,Isabella Pozzi,Roeland Nusselder,Sander Bohte
Spiking neural networks are being investigated both as biologically plausible mo
dels of neural computation and also as a potentially more efficient type of neur
al network. While convolutional spiking neural networks have been demonstrated t
o achieve near state-of-the-art performance, only one solution has been proposed
 to convert gated recurrent neural networks, so far.
Recurrent neural networks in the form of networks of gating memory cells have be
en central in state-of-the-art solutions in problem domains that involve sequenc
e recognition or generation. Here, we design an analog gated LSTM cell where its
 neurons can be substituted for efficient stochastic spiking neurons. These adap
tive spiking neurons implement an adaptive form of sigma-delta coding to convert
 internally computed analog activation values to spike-trains. For such neurons,
 we approximate the effective activation function, which resembles a sigmoid. We
 show how analog neurons with such activation functions can be used to create an
 analog LSTM cell; networks of these cells can then be trained with standard bac
kpropagation. We train these LSTM networks on a noisy and noiseless version of t
he original sequence prediction task from Hochreiter & Schmidhuber (1997), and a
lso on a noisy and noiseless version of a classical working memory reinforcement
 learning task, the T-Maze. Substituting the analog neurons for corresponding ad
aptive spiking neurons, we then show that almost all resulting spiking neural ne
twork equivalents correctly compute the original tasks.
****

Emergent Communication through Negotiation
Kris Cao,Angeliki Lazaridou,Marc Lanctot,Joel Z Leibo,Karl Tuyls,Stephen Clark
Multi-agent reinforcement learning offers a way to study how communication could
 emerge in communities of agents needing to solve specific problems. In this pap
er, we study the emergence of communication in the negotiation environment, a se
mi-cooperative model of agent interaction. We introduce two communication protoc
ols - one grounded in the semantics of the game, and one which is a priori ungro
unded. We show that self-interested agents can use the pre-grounded communicati
on channel to negotiate fairly, but are unable to effectively use the ungrounded
, cheap talk channel to do the same. However, prosocial agents do learn to use
cheap talk to find an optimal negotiating strategy, suggesting that cooperation

is necessary for language to emerge. We also study communication behaviour in a setting where one agent interacts with agents in a community with different levels of prosociality and show how agent identifiability can aid negotiation.
****
Depth separation and weight-width trade-offs for sigmoidal neural networks
Amit Deshpande,Navin Goyal,Sushrut Karmalkar
Some recent work has shown separation between the expressive power of depth-2 and depth-3 neural networks. These separation results are shown by constructing functions and input distributions, so that the function is well-approximable by a depth-3 neural network of polynomial size but it cannot be well-approximated under the chosen input distribution by any depth-2 neural network of polynomial size. These results are not robust and require carefully chosen functions as well as input distributions.

We show a similar separation between the expressive power of depth-2 and depth-3 sigmoidal neural networks over a large class of input distributions, as long as the weights are polynomially bounded. While doing so, we also show that depth-2 sigmoidal neural networks with small width and small weights can be well-approximated by low-degree multivariate polynomials.
****
Hyperedge2vec: Distributed Representations for Hyperedges
Ankit Sharma,Shafiq Joty,Himanshu Kharkwal,Jaideep Srivastava
Data structured in form of overlapping or non-overlapping sets is found in a variety of domains, sometimes explicitly but often subtly. For example, teams, which are of prime importance in social science studies are \enquote{sets of individuals}; \enquote{item sets} in pattern mining are sets; and for various types of analysis in language studies a sentence can be considered as a \enquote{set or bag of words}. Although building models and inference algorithms for structured data has been an important task in the fields of machine learning and statistics, research on \enquote{set-like} data still remains less explored. Relationships between pairs of elements can be modeled as edges in a graph. However, modeling relationships that involve all members of a set, a hyperedge is a more natural representation for the set. In this work, we focus on the problem of embedding hyperedges in a hypergraph (a network of overlapping sets) to a low dimensional vector space. We propose a probabilistic deep-learning based method as well as a tensor-based algebraic model, both of which capture the hypergraph structure in a principled manner without loosing set-level information. Our central focus is to highlight the connection between hypergraphs (topology), tensors (algebra) and probabilistic models. We present a number of interesting baselines, some of which adapt existing node-level embedding models to the hyperedge-level, as well as sequence based language techniques which are adapted for set structured hypergraph topology. The performance is evaluated with a network of social groups and a network of word phrases. Our experiments show that accuracy wise our methods perform similar to those of baselines which are not designed for hypergraphs. Moreover, our tensor based method is quiet efficient as compared to deep-learning based auto-encoder method. We therefore, argue that we have proposed more general methods which are suited for hypergraphs (and therefore also for graphs) while maintaining accuracy and efficiency.
****
A Classification-Based Perspective on GAN Distributions
Shibani Santurkar,Ludwig Schmidt,Aleksander Madry
A fundamental, and still largely unanswered, question in the context of Generative Adversarial Networks (GANs) is whether GANs are actually able to capture the key characteristics of the datasets they are trained on. The current approaches to examining this issue require significant human supervision, such as visual inspection of sampled images, and often offer only fairly limited scalability. In this paper, we propose new techniques that employ classification-based perspective to evaluate synthetic GAN distributions and their capability to accurately reflect the essential properties of the training data. These techniques require only minimal human supervision and can easily be scaled and adapted to evaluate a

variety of state-of-the-art GANs on large, popular datasets. They also indicate that GANs have significant problems in reproducing the more distributional prope rties of the training dataset. In particular, the diversity of such synthetic da ta is orders of magnitude smaller than that of the original data.
****
Multi-Scale Dense Networks for Resource Efficient Image Classification
Gao Huang,Danlu Chen,Tianhong Li,Felix Wu,Laurens van der Maaten,Kilian Weinberg er
In this paper we investigate image classification with computational resource li mits at test time. Two such settings are: 1. anytime classification, where the n etwork’s prediction for a test example is progressively updated, facilitating th e output of a prediction at any time; and 2. budgeted batch classification, wher e a fixed amount of computation is available to classify a set of examples that can be spent unevenly across “easier” and “harder” inputs. In contrast to most p rior work, such as the popular Viola and Jones algorithm, our approach is based on convolutional neural networks. We train multiple classifiers with varying res ource demands, which we adaptively apply during test time. To maximally re-use c omputation between the classifiers, we incorporate them as early-exits into a si ngle deep convolutional neural network and inter-connect them with dense connect ivity. To facilitate high quality classification early on, we use a two-dimensio nal multi-scale network architecture that maintains coarse and fine level featur es all-throughout the network. Experiments on three image-classification tasks d emonstrate that our framework substantially improves the existing state-of-the-a rt in both settings.
****
Learning to Multi-Task by Active Sampling
Sahil Sharma*,Ashutosh Kumar Jha*,Parikshit S Hegde,Balaraman Ravindran
One of the long-standing challenges in Artificial Intelligence for learning goal -directed behavior is to build a single agent which can solve multiple tasks. Re cent progress in multi-task learning for goal-directed sequential problems has b een in the form of distillation based learning wherein a student network learns from multiple task-specific expert networks by mimicking the task-specific polic ies of the expert networks. While such approaches offer a promising solution to the multi-task learning problem, they require supervision from large expert netw orks which require extensive data and computation time for training.
In this work, we propose an efficient multi-task learning framework which solves multiple goal-directed tasks in an on-line setup without the need for expert su pervision. Our work uses active learning principles to achieve multi-task learni ng by sampling the harder tasks more than the easier ones. We propose three dist inct models under our active sampling framework. An adaptive method with extreme ly competitive multi-tasking performance. A UCB-based meta-learner which casts t he problem of picking the next task to train on as a multi-armed bandit problem. A meta-learning method that casts the next-task picking problem as a full Reinf orcement Learning problem and uses actor-critic methods for optimizing the multi -tasking performance directly. We demonstrate results in the Atari 2600 domain o n seven multi-tasking instances: three 6-task instances, one 8-task instance, tw o 12-task instances and one 21-task instance.
****
Influence-Directed Explanations for Deep Convolutional Networks
Anupam Datta,Matt Fredrikson,Klas Leino,Linyi Li,Shayak Sen
We study the problem of explaining a rich class of behavioral properties of deep neural networks. Our influence-directed explanations approach this problem by p eering inside the network to identify neurons with high influence on the propert y of interest using an axiomatically justified influence measure, and then provi ding an interpretation for the concepts these neurons represent. We evaluate our approach by training convolutional neural networks on Pubfig, ImageNet, and Dia betic Retinopathy datasets. Our evaluation demonstrates that influence-directed explanations (1) localize features used by the network, (2) isolate features di stinguishing related instances, (3) help extract the essence of what the network learned about the class, and (4) assist in debugging misclassifications.

****
Variational Network Quantization
Jan Achterhold,Jan Mathias Koehler,Anke Schmeink,Tim Genewein
In this paper, the preparation of a neural network for pruning and few-bit quant
ization is formulated as a variational inference problem. To this end, a quantiz
ing prior that leads to a multi-modal, sparse posterior distribution over weight
s, is introduced and a differentiable Kullback-Leibler divergence approximation
for this prior is derived. After training with Variational Network Quantization,
 weights can be replaced by deterministic quantization values with small to negl
igible loss of task accuracy (including pruning by setting weights to 0). The me
thod does not require fine-tuning after quantization. Results are shown for tern
ary quantization on LeNet-5 (MNIST) and DenseNet (CIFAR-10).
****
Variational Continual Learning
Cuong V. Nguyen,Yingzhen Li,Thang D. Bui,Richard E. Turner
This paper develops variational continual learning (VCL), a simple but general f
ramework for continual learning that fuses online variational inference (VI) and
 recent advances in Monte Carlo VI for neural networks. The framework can succes
sfully train both deep discriminative models and deep generative models in compl
ex continual learning settings where existing tasks evolve over time and entirel
y new tasks emerge. Experimental results show that VCL outperforms state-of-the-
art continual learning methods on a variety of tasks, avoiding catastrophic forg
etting in a fully automatic way.
****
Long Term Memory Network for Combinatorial Optimization Problems
Hazem A. A. Nomer,Abdallah Aboutahoun,Ashraf Elsayed
This paper introduces a framework for solving combinatorial optimization problem
s by learning from input-output examples of optimization problems. We introduce
a new memory augmented neural model in which the memory is not resettable (i.e t
he information stored in the memory after processing an input example is kept fo
r the next seen examples). We used deep reinforcement learning to train a memory
 controller agent to store useful memories. Our model was able to outperform han
d-crafted solver on Binary Linear Programming (Binary LP). The proposed model is
 tested on different Binary LP instances with large number of variables (up to 1
000 variables) and constrains (up to 700 constrains).
****
Evaluating the Robustness of Neural Networks: An Extreme Value Theory Approach
Tsui-Wei Weng*,Huan Zhang*,Pin-Yu Chen,Jinfeng Yi,Dong Su,Yupeng Gao,Cho-Jui Hsi
eh,Luca Daniel
The robustness of neural networks to adversarial examples has received great att
ention due to security implications. Despite various attack approaches to crafti
ng visually imperceptible adversarial examples, little has been developed toward
s a comprehensive measure of robustness. In this paper, we provide theoretical j
ustification for converting robustness analysis into a local Lipschitz constant
estimation problem, and propose to use the Extreme Value Theory for efficient ev
aluation. Our analysis yields a novel robustness metric called CLEVER, which is
short for Cross Lipschitz Extreme Value for nEtwork Robustness. The proposed CLE
VER score is attack-agnostic and is computationally feasible for large neural ne
tworks. Experimental results on various networks, including ResNet, Inception-v3
 and MobileNet, show that (i) CLEVER is aligned with the robustness indication m
easured by the ℓ_2 and ℓ_∞ norms of adversarial examples from pow
erful attacks, and (ii) defended networks using defensive distillation or bounde
d ReLU indeed give better CLEVER scores. To the best of our knowledge, CLEVER is
 the first attack-independent robustness metric that can be applied to any neura
l network classifiers.

****
Federated Learning: Strategies for Improving Communication Efficiency

Jakub Kone■ný,H. Brendan McMahan,Felix X. Yu,Ananda Theertha Suresh,Dave Bacon,P
eter Richtárik
Federated Learning is a machine learning setting where the goal is to train a hi
gh-quality centralized model while training data remains distributed over a larg
e number of clients each with unreliable and relatively slow network connections
. We consider learning algorithms for this setting where on each round, each cl
ient independently computes an update to the current model based on its local da
ta, and communicates this update to a central server, where the client-side upda
tes are aggregated to compute a new global model. The typical clients in this s
etting are mobile phones, and communication efficiency is of the utmost importan
ce.

In this paper, we propose two ways to reduce the uplink communication costs: str
uctured updates, where we directly learn an update from a restricted space param
etrized using a smaller number of variables, e.g. either low-rank or a random ma
sk; and sketched updates, where we learn a full model update and then compress i
t using a combination of quantization, random rotations, and subsampling before
sending it to the server. Experiments on both convolutional and recurrent networ
ks show that the proposed methods can reduce the communication cost by two order
s of magnitude.
****
Learning to Treat Sepsis with Multi-Output Gaussian Process Deep Recurrent Q-Net
works
Joseph Futoma,Anthony Lin,Mark Sendak,Armando Bedoya,Meredith Clement,Cara O'Bri
en,Katherine Heller
Sepsis is a life-threatening complication from infection and a leading cause of
mortality in hospitals. While early detection of sepsis improves patient outcom
es, there is little consensus on exact treatment guidelines, and treating septic
patients remains an open problem. In this work we present a new deep reinforc
ement learning method that we use to learn optimal personalized treatment polici
es for septic patients. We model patient continuous-valued physiological time se
ries using multi-output Gaussian processes, a probabilistic model that easily ha
ndles missing values and irregularly spaced observation times while maintaining
estimates of uncertainty. The Gaussian process is directly tied to a deep recurr
ent Q-network that learns clinically interpretable treatment policies, and both
models are learned together end-to-end. We evaluate our approach on a heterogen
eous dataset of septic spanning 15 months from our university health system, and
find that our learned policy could reduce patient mortality by as much as 8.2\%
from an overall baseline mortality rate of 13.3\%. Our algorithm could be used
to make treatment recommendations to physicians as part of a decision support t
ool, and the framework readily applies to other reinforcement learning problems
that rely on sparsely sampled and frequently missing multivariate time series da
ta.

****
TD or not TD: Analyzing the Role of Temporal Differencing in Deep Reinforcement
Learning
Artemij Amiranashvili,Alexey Dosovitskiy,Vladlen Koltun,Thomas Brox
Our understanding of reinforcement learning (RL) has been shaped by theoretical
and empirical results that were obtained decades ago using tabular representatio
ns and linear function approximators. These results suggest that RL methods that
use temporal differencing (TD) are superior to direct Monte Carlo estimation (M
C). How do these results hold up in deep RL, which deals with perceptually compl
ex environments and deep nonlinear models? In this paper, we re-examine the role
of TD in modern deep RL, using specially designed environments that control for
specific factors that affect performance, such as reward sparsity, reward delay
, and the perceptual complexity of the task. When comparing TD with infinite-hor
izon MC, we are able to reproduce classic results in modern settings. Yet we als
o find that finite-horizon MC is not inferior to TD, even when rewards are spars
e or delayed. This makes MC a viable alternative to TD in deep RL.

****
Model Distillation with Knowledge Transfer from Face Classification to Alignment and Verification
Chong Wang,Xipeng Lan,Yangang Zhang
Knowledge distillation is a potential solution for model compression. The idea is to make a small student network imitate the target of a large teacher network, then the student network can be competitive to the teacher one. Most previous studies focus on model distillation in the classification task, where they propose different architectures and initializations for the student network. However, only the classification task is not enough, and other related tasks such as regression and retrieval are barely considered. To solve the problem, in this paper, we take face recognition as a breaking point and propose model distillation with knowledge transfer from face classification to alignment and verification. By selecting appropriate initializations and targets in the knowledge transfer, the distillation can be easier in non-classification tasks. Experiments on the CelebA and CASIA-WebFace datasets demonstrate that the student network can be competitive to the teacher one in alignment and verification, and even surpasses the teacher network under specific compression rates. In addition, to achieve stronger knowledge transfer, we also use a common initialization trick to improve the distillation performance of classification. Evaluations on the CASIA-Webface and large-scale MS-Celeb-1M datasets show the effectiveness of this simple trick.
****
Not-So-CLEVR: Visual Relations Strain Feedforward Neural Networks
Junkyung Kim,Matthew Ricci,Thomas Serre
The robust and efficient recognition of visual relations in images is a hallmark of biological vision. Here, we argue that, despite recent progress in visual recognition, modern machine vision algorithms are severely limited in their ability to learn visual relations. Through controlled experiments, we demonstrate that visual-relation problems strain convolutional neural networks (CNNs). The networks eventually break altogether when rote memorization becomes impossible such as when the intra-class variability exceeds their capacity. We further show that another type of feedforward network, called a relational network (RN), which was shown to successfully solve seemingly difficult visual question answering (VQA) problems on the CLEVR datasets, suffers similar limitations. Motivated by the comparable success of biological vision, we argue that feedback mechanisms including working memory and attention are the key computational components underlying abstract visual reasoning.
****
A Bayesian Perspective on Generalization and Stochastic Gradient Descent
Samuel L. Smith and Quoc V. Le
We consider two questions at the heart of machine learning; how can we predict if a minimum will generalize to the test set, and why does stochastic gradient descent find minima that generalize well? Our work responds to \citet{zhang2016understanding}, who showed deep neural networks can easily memorize randomly labeled training data, despite generalizing well on real labels of the same inputs. We show that the same phenomenon occurs in small linear models. These observations are explained by the Bayesian evidence, which penalizes sharp minima but is invariant to model parameterization. We also demonstrate that, when one holds the learning rate fixed, there is an optimum batch size which maximizes the test set accuracy. We propose that the noise introduced by small mini-batches drives the parameters towards minima whose evidence is large. Interpreting stochastic gradient descent as a stochastic differential equation, we identify the ``noise scale" $g = \epsilon (\frac{N}{B} - 1) \approx \epsilon N/B$, where ϵ is the learning rate, N the training set size and B the batch size. Consequently the optimum batch size is proportional to both the learning rate and the size of the training set, $B_{opt} \propto \epsilon N$. We verify these predictions empirically.
****
Learning Non-Metric Visual Similarity for Image Retrieval
Noa Garcia,George Vogiatzis

Measuring visual (dis)similarity between two or more instances within a data dis
tribution is a fundamental task in many applications, specially in image retriev
al. Theoretically, non-metric distances are able to generate a more complex and
accurate similarity model than metric distances, provided that the non-linear da
ta distribution is precisely captured by the similarity model. In this work, we
analyze a simple approach for deep learning networks to be used as an approximat
ion of non-metric similarity functions and we study how these models generalize
across different image retrieval datasets.
****
AmbientGAN: Generative models from lossy measurements
Ashish Bora,Eric Price,Alexandros G. Dimakis
Generative models provide a way to model structure in complex distributions and
have been shown to be useful for many tasks of practical interest. However, curr
ent techniques for training generative models require access to fully-observed s
amples. In many settings, it is expensive or even impossible to obtain fully-obs
erved samples, but economical to obtain partial, noisy observations. We consider
 the task of learning an implicit generative model given only lossy measurements
 of samples from the distribution of interest. We show that the true underlying
distribution can be provably recovered even in the presence of per-sample inform
ation loss for a class of measurement models. Based on this, we propose a new me
thod of training Generative Adversarial Networks (GANs) which we call AmbientGAN
. On three benchmark datasets, and for various measurement models, we demonstrat
e substantial qualitative and quantitative improvements. Generative models train
ed with our method can obtain 2-4x higher inception scores than the baseline
s.
****
A Painless Attention Mechanism for Convolutional Neural Networks
Pau Rodríguez,Guillem Cucurull,Jordi Gonzàlez,Josep M. Gonfaus,Xavier Roca
We propose a novel attention mechanism to enhance Convolutional Neural Networks
for fine-grained recognition. The proposed mechanism reuses CNN feature activati
ons to find the most informative parts of the image at different depths with the
 help of gating mechanisms and without part annotations. Thus, it can be used to
 augment any layer of a CNN to extract low- and high-level local information to
be more discriminative.

Differently, from other approaches, the mechanism we propose just needs a single
 pass through the input and it can be trained end-to-end through SGD. As a conse
quence, the proposed mechanism is modular, architecture-independent, easy to imp
lement, and faster than iterative approaches.

Experiments show that, when augmented with our approach, Wide Residual Networks
systematically achieve superior performance on each of five different fine-grain
ed recognition datasets: the Adience age and gender recognition benchmark, Calte
ch-UCSD Birds-200-2011, Stanford Dogs, Stanford Cars, and UEC Food-100, obtainin
g competitive and state-of-the-art scores.
****
Anomaly Detection with Generative Adversarial Networks
Lucas Deecke,Robert Vandermeulen,Lukas Ruff,Stephan Mandt,Marius Kloft
Many anomaly detection methods exist that perform well on low-dimensional proble
ms however there is a notable lack of effective methods for high-dimensional spa
ces, such as images. Inspired by recent successes in deep learning we propose a
novel approach to anomaly detection using generative adversarial networks. Given
 a sample under consideration, our method is based on searching for a good repre
sentation of that sample in the latent space of the generator; if such a represe
ntation is not found, the sample is deemed anomalous. We achieve state-of-the-a
rt performance on standard image benchmark datasets and visual inspection of the
 most anomalous samples reveals that our method does indeed return anomalies.
****
A Spectral Approach to Generalization and Optimization in Neural Networks
Farzan Farnia,Jesse Zhang,David Tse

The recent success of deep neural networks stems from their ability to generalize well on real data; however, Zhang et al. have observed that neural networks can easily overfit random labels. This observation demonstrates that with the existing theory, we cannot adequately explain why gradient methods can find generalizable solutions for neural networks. In this work, we use a Fourier-based approach to study the generalization properties of gradient-based methods over 2-layer neural networks with sinusoidal activation functions. We prove that if the underlying distribution of data has nice spectral properties such as bandlimitedness, then the gradient descent method will converge to generalizable local minima. We also establish a Fourier-based generalization bound for bandlimited spaces, which generalizes to other activation functions. Our generalization bound motivates a grouped version of path norms for measuring the complexity of 2-layer neural networks with ReLU activation functions. We demonstrate numerically that regularization of this group path norm results in neural network solutions that can fit true labels without losing test accuracy while not overfitting random labels.
****
Fix your classifier: the marginal value of training the last weight layer
Elad Hoffer,Itay Hubara,Daniel Soudry
Neural networks are commonly used as models for classification for a wide variety of tasks. Typically, a learned affine transformation is placed at the end of such models, yielding a per-class value used for classification. This classifier can have a vast number of parameters, which grows linearly with the number of possible classes, thus requiring increasingly more resources.

In this work we argue that this classifier can be fixed, up to a global scale constant, with little or no loss of accuracy for most tasks, allowing memory and computational benefits. Moreover, we show that by initializing the classifier with a Hadamard matrix we can speed up inference as well. We discuss the implications for current understanding of neural network models.

****
Link Weight Prediction with Node Embeddings
Yuchen Hou,Lawrence B. Holder
Application of deep learning has been successful in various domains such as image recognition, speech recognition and natural language processing. However, the research on its application in graph mining is still in an early stage. Here we
present the first generic deep learning approach to the graph link weight prediction
problem based on node embeddings. We evaluate this approach with three different node embedding techniques experimentally and compare its performance with two state-of-the-art non deep learning baseline approaches. Our experiment results suggest that this deep learning approach outperforms the baselines by up to
70% depending on the dataset and embedding technique applied. This approach shows that deep learning can be successfully applied to link weight prediction to
improve prediction accuracy.
****
Compositional Obverter Communication Learning from Raw Visual Input
Edward Choi,Angeliki Lazaridou,Nando de Freitas
One of the distinguishing aspects of human language is its compositionality, which allows us to describe complex environments with limited vocabulary. Previously, it has been shown that neural network agents can learn to communicate in a highly structured, possibly compositional language based on disentangled input (e.g. hand- engineered features). Humans, however, do not learn to communicate based on well-summarized features. In this work, we train neural agents to simultaneously develop visual perception from raw image pixels, and learn to communicate with a sequence of discrete symbols. The agents play an image description game where the image contains factors such as colors and shapes. We train the agents u

sing the obverter technique where an agent introspects to generate messages that maximize its own understanding. Through qualitative analysis, visualization and a zero-shot test, we show that the agents can develop, out of raw image pixels, a language with compositional properties, given a proper pressure from the environment.
****
Temporally Efficient Deep Learning with Spikes
Peter O'Connor,Efstratios Gavves,Matthias Reisser,Max Welling
The vast majority of natural sensory data is temporally redundant. For instance, video frames or audio samples which are sampled at nearby points in time tend to have similar values. Typically, deep learning algorithms take no advantage of this redundancy to reduce computations. This can be an obscene waste of energy. We present a variant on backpropagation for neural networks in which computation scales with the rate of change of the data - not the rate at which we process the data. We do this by implementing a form of Predictive Coding wherein neurons communicate a combination of their state, and their temporal change in state, and quantize this signal using Sigma-Delta modulation. Intriguingly, this simple communication rule give rise to units that resemble biologically-inspired leaky integrate-and-fire neurons, and to a spike-timing-dependent weight-update similar to Spike-Timing Dependent Plasticity (STDP), a synaptic learning rule observed in the brain. We demonstrate that on MNIST, on a temporal variant of MNIST, and on Youtube-BB, a dataset with videos in the wild, our algorithm performs about as well as a standard deep network trained with backpropagation, despite only communicating discrete values between layers.
****
Neural Program Search: Solving Data Processing Tasks from Description and Examples
Illia Polosukhin,Alexander Skidanov
We present a Neural Program Search, an algorithm to generate programs from natural language description and a small number of input / output examples. The algorithm combines methods from Deep Learning and Program Synthesis fields by designing rich domain-specific language (DSL) and defining efficient search algorithm guided by a Seq2Tree model on it. To evaluate the quality of the approach we also present a semi-synthetic dataset of descriptions with test examples and corresponding programs. We show that our algorithm significantly outperforms sequence-to-sequence model with attention baseline.
****
End-to-End Abnormality Detection in Medical Imaging
Dufan Wu,Kyungsang Kim,Bin Dong,Quanzheng Li
Deep neural networks (DNN) have shown promising performance in computer vision. In medical imaging, encouraging results have been achieved with deep learning for applications such as segmentation, lesion detection and classification. Nearly all of the deep learning based image analysis methods work on reconstructed images, which are obtained from original acquisitions via solving inverse problems (reconstruction). The reconstruction algorithms are designed for human observers, but not necessarily optimized for DNNs which can often observe features that are incomprehensible for human eyes. Hence, it is desirable to train the DNNs directly from the original data which lie in a different domain with the images. In this paper, we proposed an end-to-end DNN for abnormality detection in medical imaging. To align the acquisition with the annotations made by radiologists in the image domain, a DNN was built as the unrolled version of iterative reconstruction algorithms to map the acquisitions to images, and followed by a 3D convolutional neural network (CNN) to detect the abnormality in the reconstructed images. The two networks were trained jointly in order to optimize the entire DNN for the detection task from the original acquisitions. The DNN was implemented for lung nodule detection in low-dose chest computed tomography (CT), where a numerical simulation was done to generate acquisitions from 1,018 chest CT images with radiologists' annotations. The proposed end-to-end DNN demonstrated better sensitivity and accuracy for the task compared to a two-step approach, in which the reconstruction and detection DNNs were trained separately. A significant reductio

n of false positive rate on suspicious lesions were observed, which is crucial for the known over-diagnosis in low-dose lung CT imaging. The images reconstructed by the proposed end-to-end network also presented enhanced details in the region of interest.
****
Transformation Autoregressive Networks
Junier Oliva,Avinava Dubey,Barnabás Póczos,Eric P. Xing,Jeff Schneider
The fundamental task of general density estimation has been of keen interest to machine learning. Recent advances in density estimation have either: a) proposed using a flexible model to estimate the conditional factors of the chain rule; or b) used flexible, non-linear transformations of variables of a simple base distribution. Instead, this work jointly leverages transformations of variables and autoregressive conditional models, and proposes novel methods for both. We provide a deeper understanding of our models, showing a considerable improvement with our methods through a comprehensive study over both real world and synthetic data. Moreover, we illustrate the use of our models in outlier detection and image modeling task.
****
Countering Adversarial Images using Input Transformations
Chuan Guo,Mayank Rana,Moustapha Cisse,Laurens van der Maaten
This paper investigates strategies that defend against adversarial-example attacks on image-classification systems by transforming the inputs before feeding them to the system. Specifically, we study applying image transformations such as bit-depth reduction, JPEG compression, total variance minimization, and image quilting before feeding the image to a convolutional network classifier. Our experiments on ImageNet show that total variance minimization and image quilting are very effective defenses in practice, in particular, when the network is trained on transformed images. The strength of those defenses lies in their non-differentiable nature and their inherent randomness, which makes it difficult for an adversary to circumvent the defenses. Our best defense eliminates 60% of strong gray-box and 90% of strong black-box attacks by a variety of major attack methods.
****
Generalization of Learning using Reservoir Computing
Sanjukta Krishnagopal,Yiannis Aloimonos,Michelle Girvan
We investigate the methods by which a Reservoir Computing Network (RCN) learns concepts such as 'similar' and 'different' between pairs of images using a small training dataset and generalizes these concepts to previously unseen types of data. Specifically, we show that an RCN trained to identify relationships between image-pairs drawn from a subset of digits from the MNIST database or the depth maps of subset of visual scenes from a moving camera generalizes the learned transformations to images of digits unseen during training or depth maps of different visual scenes. We infer, using Principal Component Analysis, that the high dimensional reservoir states generated from an input image pair with a specific transformation converge over time to a unique relationship. Thus, as opposed to training the entire high dimensional reservoir state, the reservoir only needs to train on these unique relationships, allowing the reservoir to perform well with very few training examples. Thus, generalization of learning to unseen images is interpretable in terms of clustering of the reservoir state onto the attractor corresponding to the transformation in reservoir space. We find that RCNs can identify and generalize linear and non-linear transformations, and combinations of transformations, naturally and be a robust and effective image classifier. Additionally, RCNs perform significantly better than state of the art neural network classification techniques such as deep Siamese Neural Networks (SNNs) in generalization tasks both on the MNIST dataset and more complex depth maps of visual scenes from a moving camera. This work helps bridge the gap between explainable machine learning and biological learning through analogies using small datasets, and points to new directions in the investigation of learning processes.
****
Fraternal Dropout
Konrad Zolna,Devansh Arpit,Dendi Suhubdy,Yoshua Bengio

Recurrent neural networks (RNNs) are important class of architectures among neural networks useful for language modeling and sequential prediction. However, optimizing RNNs is known to be harder compared to feed-forward neural networks. A number of techniques have been proposed in literature to address this problem. In this paper we propose a simple technique called fraternal dropout that takes advantage of dropout to achieve this goal. Specifically, we propose to train two identical copies of an RNN (that share parameters) with different dropout masks while minimizing the difference between their (pre-softmax) predictions. In this way our regularization encourages the representations of RNNs to be invariant to dropout mask, thus being robust. We show that our regularization term is upper bounded by the expectation-linear dropout objective which has been shown to address the gap due to the difference between the train and inference phases of dropout. We evaluate our model and achieve state-of-the-art results in sequence modeling tasks on two benchmark datasets - Penn Treebank and Wikitext-2. We also show that our approach leads to performance improvement by a significant margin in image captioning (Microsoft COCO) and semi-supervised (CIFAR-10) tasks.
****

MGAN: Training Generative Adversarial Nets with Multiple Generators

Quan Hoang,Tu Dinh Nguyen,Trung Le,Dinh Phung

We propose in this paper a new approach to train the Generative Adversarial Nets (GANs) with a mixture of generators to overcome the mode collapsing problem. The main intuition is to employ multiple generators, instead of using a single one as in the original GAN. The idea is simple, yet proven to be extremely effective at covering diverse data modes, easily overcoming the mode collapsing problem and delivering state-of-the-art results. A minimax formulation was able to establish among a classifier, a discriminator, and a set of generators in a similar spirit with GAN. Generators create samples that are intended to come from the same distribution as the training data, whilst the discriminator determines whether samples are true data or generated by generators, and the classifier specifies which generator a sample comes from. The distinguishing feature is that internal samples are created from multiple generators, and then one of them will be randomly selected as final output similar to the mechanism of a probabilistic mixture model. We term our method Mixture Generative Adversarial Nets (MGAN). We develop theoretical analysis to prove that, at the equilibrium, the Jensen-Shannon divergence (JSD) between the mixture of generators' distributions and the empirical data distribution is minimal, whilst the JSD among generators' distributions is maximal, hence effectively avoiding the mode collapsing problem. By utilizing parameter sharing, our proposed model adds minimal computational cost to the standard GAN, and thus can also efficiently scale to large-scale datasets. We conduct extensive experiments on synthetic 2D data and natural image databases (CIFAR-10, STL-10 and ImageNet) to demonstrate the superior performance of our MGAN in achieving state-of-the-art Inception scores over latest baselines, generating diverse and appealing recognizable objects at different resolutions, and specializing in capturing different types of objects by the generators.
****

Exploring Sentence Vectors Through Automatic Summarization

Adly Templeton,Jugal Kalita

Vector semantics, especially sentence vectors, have recently been used successfully in many areas of natural language processing. However, relatively little work has explored the internal structure and properties of spaces of sentence vectors. In this paper, we will explore the properties of sentence vectors by studying a particular real-world application: Automatic Summarization. In particular, we show that cosine similarity between sentence vectors and document vectors is strongly correlated with sentence importance and that vector semantics can identify and correct gaps between the sentences chosen so far and the document. In addition, we identify specific dimensions which are linked to effective summaries. To our knowledge, this is the first time specific dimensions of sentence embeddings have been connected to sentence properties. We also compare the features of different methods of sentence embeddings. Many of these insights have applications in uses of sentence embeddings far beyond summarization.

****
Noise-Based Regularizers for Recurrent Neural Networks
Adji B. Dieng,Jaan Altosaar,Rajesh Ranganath,David M. Blei
Recurrent neural networks (RNNs) are powerful models for sequential data. They can approximate arbitrary computations, and have been used successfully in domains such as text and speech. However, the flexibility of RNNs makes them susceptible to overfitting and regularization is important. We develop a noise-based regularization method for RNNs. The idea is simple and easy to implement: we inject noise in the hidden units of the RNN and then maximize the original RNN's likelihood averaged over the injected noise. On a language modeling benchmark, our method achieves better performance than the deterministic RNN and the variational dropout.
****
Generative Models of Visually Grounded Imagination
Ramakrishna Vedantam,Ian Fischer,Jonathan Huang,Kevin Murphy
It is easy for people to imagine what a man with pink hair looks like, even if they have never seen such a person before. We call the ability to create images of novel semantic concepts visually grounded imagination. In this paper, we show how we can modify variational auto-encoders to perform this task. Our method uses a novel training objective, and a novel product-of-experts inference network, which can handle partially specified (abstract) concepts in a principled and efficient way. We also propose a set of easy-to-compute evaluation metrics that capture our intuitive notions of what it means to have good visual imagination, namely correctness, coverage, and compositionality (the 3 C’s). Finally, we perform a detailed comparison of our method with two existing joint image-attribute VAE methods (the JMVAE method of Suzuki et al., 2017 and the BiVCCA method of Wang et al., 2016) by applying them to two datasets: the MNIST-with-attributes dataset (which we introduce here), and the CelebA dataset (Liu et al., 2015).
****
Unsupervised Hierarchical Video Prediction
Nevan Wichers,Dumitru Erhan,Honglak Lee
Much recent research has been devoted to video prediction and generation, but mostly for short-scale time horizons. The hierarchical video prediction method by Villegas et al. (2017) is an example of a state of the art method for long term video prediction. However, their method has limited applicability in practical settings as it requires a ground truth pose (e.g., poses of joints of a human) at training time.   This paper presents a long term hierarchical video prediction model that does not have such a restriction. We show that the network learns its own higher-level structure (e.g., pose equivalent hidden variables) that works better in cases where the ground truth pose does not fully capture all of the information needed to predict the next frame. This method gives sharper results than other video prediction methods which do not require a ground truth pose, and its efficiency is shown on the Humans 3.6M and Robot Pushing datasets.
****
Sparse-Complementary Convolution for Efficient Model Utilization on CNNs
Chun-Fu (Richard) Chen,Jinwook Oh,Quanfu Fan,Marco Pistoia,Gwo Giun (Chris) Lee
We introduce an efficient way to increase the accuracy of convolution neural networks (CNNs) based on high model utilization without increasing any computational complexity.
The proposed sparse-complementary convolution replaces regular convolution with sparse and complementary shapes of kernels, covering the same receptive field.
By the nature of deep learning, high model utilization of a CNN can be achieved with more simpler kernels rather than fewer complex kernels.
This simple but insightful model reuses of recent network architectures, ResNet and DenseNet, can provide better accuracy for most classification tasks (CIFAR-10/100 and ImageNet) compared to their baseline models. By simply replacing the convolution of a CNN with our sparse-complementary convolution, at the same FLOPs and parameters, we can improve top-1 accuracy on ImageNet by 0.33% and 0.18% for ResNet-101 and ResNet-152, respectively. A similar accuracy improvement could

be gained by increasing the number of layers in those networks by ~1.5x.
****
An empirical study on evaluation metrics of generative adversarial networks
Gao Huang,Yang Yuan,Qiantong Xu,Chuan Guo,Yu Sun,Felix Wu,Kilian Weinberger
Despite the widespread interest in generative adversarial networks (GANs), few works have studied the metrics that quantitatively evaluate GANs' performance. In this paper, we revisit several representative sample-based evaluation metrics for GANs, and address the important problem of \emph{how to evaluate the evaluation metrics}. We start with a few necessary conditions for metrics to produce meaningful scores, such as distinguishing real from generated samples, identifying mode dropping and mode collapsing, and detecting overfitting. Then with a series of carefully designed experiments, we are able to comprehensively investigate existing sample-based metrics and identify their strengths and limitations in practical settings. Based on these results, we observe that kernel Maximum Mean Discrepancy (MMD) and the 1-Nearest-Neighbour (1-NN) two-sample test seem to satisfy most of the desirable properties, provided that the distances between samples are computed in a suitable feature space. Our experiments also unveil interesting properties about the behavior of several popular GAN models, such as whether they are memorizing training samples, and how far these state-of-the-art GANs are from perfect.
****
Soft Value Iteration Networks for Planetary Rover Path Planning
Max Pflueger,Ali Agha,Gaurav S. Sukhatme
Value iteration networks are an approximation of the value iteration (VI) algorithm implemented with convolutional neural networks to make VI fully differentiable. In this work, we study these networks in the context of robot motion planning, with a focus on applications to planetary rovers. The key challenging task in learning-based motion planning is to learn a transformation from terrain observations to a suitable navigation reward function. In order to deal with complex terrain observations and policy learning, we propose a value iteration recurrence, referred to as the soft value iteration network (SVIN). SVIN is designed to produce more effective training gradients through the value iteration network. It relies on a soft policy model, where the policy is represented with a probability distribution over all possible actions, rather than a deterministic policy that returns only the best action. We demonstrate the effectiveness of the proposed method in robot motion planning scenarios. In particular, we study the application of SVIN to very challenging problems in planetary rover navigation and present early training results on data gathered by the Curiosity rover that is currently operating on Mars.
****
Discovering Order in Unordered Datasets: Generative Markov Networks
Yao-Hung Hubert Tsai,Han Zhao,Nebojsa Jojic,Ruslan Salakhutdinov
The assumption that data samples are independently identically distributed is the backbone of many learning algorithms. Nevertheless, datasets often exhibit rich structures in practice, and we argue that there exist some unknown orders within the data instances. Aiming to find such orders, we introduce a novel Generative Markov Network (GMN) which we use to extract the order of data instances automatically. Specifically, we assume that the instances are sampled from a Markov chain. Our goal is to learn the transitional operator of the chain as well as the generation order by maximizing the generation probability under all possible data permutations. One of our key ideas is to use neural networks as a soft lookup table for approximating the possibly huge, but discrete transition matrix. This strategy allows us to amortize the space complexity with a single model and make the transitional operator generalizable to unseen instances. To ensure the learned Markov chain is ergodic, we propose a greedy batch-wise permutation scheme that allows fast training. Empirically, we evaluate the learned Markov chain by showing that GMNs are able to discover orders among data instances and also perform comparably well to state-of-the-art methods on the one-shot recognition benchmark task.
****

Efficiently applying attention to sequential data with the Recurrent Discounted Attention unit
Brendan Maginnis,Pierre Richemond
Recurrent Neural Networks architectures excel at processing sequences by modelling dependencies over different timescales. The recently introduced Recurrent Weighted Average (RWA) unit captures long term dependencies far better than an LSTM on several challenging tasks. The RWA achieves this by applying attention to each input and computing a weighted average over the full history of its computations. Unfortunately, the RWA cannot change the attention it has assigned to previous timesteps, and so struggles with carrying out consecutive tasks or tasks with changing requirements.
We present the Recurrent Discounted Attention (RDA) unit that builds on the RWA by additionally allowing the discounting of the past.
We empirically compare our model to RWA, LSTM and GRU units on several challenging tasks. On tasks with a single output the RWA, RDA and GRU units learn much quicker than the LSTM and with better performance.
On the multiple sequence copy task our RDA unit learns the task three times as quickly as the LSTM or GRU units while the RWA fails to learn at all. On the Wikipedia character prediction task the LSTM performs best but it followed closely by our RDA unit. Overall our RDA unit performs well and is sample efficient on a large variety of sequence tasks.
****
Beyond Word Importance: Contextual Decomposition to Extract Interactions from LSTMs
W. James Murdoch,Peter J. Liu,Bin Yu
The driving force behind the recent success of LSTMs has been their ability to learn complex and non-linear relationships. Consequently, our inability to describe these relationships has led to LSTMs being characterized as black boxes. To this end, we introduce contextual decomposition (CD), an interpretation algorithm for analysing individual predictions made by standard LSTMs, without any changes to the underlying model. By decomposing the output of a LSTM, CD captures the contributions of combinations of words or variables to the final prediction of an LSTM. On the task of sentiment analysis with the Yelp and SST data sets, we show that CD is able to reliably identify words and phrases of contrasting sentiment, and how they are combined to yield the LSTM's final prediction. Using the phrase-level labels in SST, we also demonstrate that CD is able to successfully extract positive and negative negations from an LSTM, something which has not previously been done.
****
Multi-View Data Generation Without View Supervision
Mickael Chen,Ludovic Denoyer,Thierry Artières
The development of high-dimensional generative models has recently gained a great surge of interest with the introduction of variational auto-encoders and generative adversarial neural networks. Different variants have been proposed where the underlying latent space is structured, for example, based on attributes describing the data to generate. We focus on a particular problem where one aims at generating samples corresponding to a number of objects under various views. We assume that the distribution of the data is driven by two independent latent factors: the content, which represents the intrinsic features of an object, and the view, which stands for the settings of a particular observation of that object. Therefore, we propose a generative model and a conditional variant built on such a disentangled latent space. This approach allows us to generate realistic samples corresponding to various objects in a high variety of views. Unlike many multi-view approaches, our model doesn't need any supervision on the views but only on the content. Compared to other conditional generation approaches that are mostly based on binary or categorical attributes, we make no such assumption about the factors of variations. Our model can be used on problems with a huge, potentially infinite, number of categories. We experiment it on four images datasets on which we demonstrate the effectiveness of the model and its ability to generalize.

****
Multi-level Residual Networks from Dynamical Systems View
Bo Chang,Lili Meng,Eldad Haber,Frederick Tung,David Begert
Deep residual networks (ResNets) and their variants are widely used in many computer vision applications and natural language processing tasks. However, the theoretical principles for designing and training ResNets are still not fully understood. Recently, several points of view have emerged to try to interpret ResNet theoretically, such as unraveled view, unrolled iterative estimation and dynamical systems view. In this paper, we adopt the dynamical systems point of view, and analyze the lesioning properties of ResNet both theoretically and experimentally. Based on these analyses, we additionally propose a novel method for accelerating ResNet training. We apply the proposed method to train ResNets and Wide ResNets for three image classification benchmarks, reducing training time by more than 40\% with superior or on-par accuracy.
****
Deep Learning for Physical Processes: Incorporating Prior Scientific Knowledge
Emmanuel de Bezenac,Arthur Pajot,Patrick Gallinari
We consider the use of Deep Learning methods for modeling complex phenomena like those occurring in natural physical processes. With the large amount of data gathered on these phenomena the data intensive paradigm could begin to challenge more traditional approaches elaborated over the years in fields like maths or physics. However, despite considerable successes in a variety of application domains, the machine learning field is not yet ready to handle the level of complexity required by such problems. Using an example application, namely Sea Surface Temperature Prediction, we show how general background knowledge gained from the physics could be used as a guideline for designing efficient Deep Learning models. In order to motivate the approach and to assess its generality we demonstrate a formal link between the solution of a class of differential equations underlying a large family of physical phenomena and the proposed model. Experiments and comparison with series of baselines including a state of the art numerical approach is then provided.
****
Natural Language Inference over Interaction Space
Yichen Gong,Heng Luo,Jian Zhang
Natural Language Inference (NLI) task requires an agent to determine the logical relationship between a natural language premise and a natural language hypothesis. We introduce Interactive Inference Network (IIN), a novel class of neural network architectures that is able to achieve high-level understanding of the sentence pair by hierarchically extracting semantic features from interaction space. We show that an interaction tensor (attention weight) contains semantic information to solve natural language inference, and a denser interaction tensor contains richer semantic information. One instance of such architecture, Densely Interactive Inference Network (DIIN), demonstrates the state-of-the-art performance on large scale NLI copora and large-scale NLI alike corpus. It's noteworthy that DIIN achieve a greater than 20% error reduction on the challenging Multi-Genre NLI (MultiNLI) dataset with respect to the strongest published system.
****
Wasserstein Auto-Encoders
Ilya Tolstikhin,Olivier Bousquet,Sylvain Gelly,Bernhard Schoelkopf
We propose the Wasserstein Auto-Encoder (WAE)---a new algorithm for building a generative model of the data distribution. WAE minimizes a penalized form of the Wasserstein distance between the model distribution and the target distribution, which leads to a different regularizer than the one used by the Variational Auto-Encoder (VAE).
This regularizer encourages the encoded training distribution to match the prior. We compare our algorithm with several other techniques and show that it is a generalization of adversarial auto-encoders (AAE). Our experiments show that WAE shares many of the properties of VAEs (stable training, encoder-decoder architecture, nice latent manifold structure) while generating samples of better quality.

****
Make SVM great again with Siamese kernel for few-shot learning
Bence Tilk
While deep neural networks have shown outstanding results in a wide range of applications,
learning from a very limited number of examples is still a challenging
task. Despite the difficulties of the few-shot learning, metric-learning techniques
showed the potential of the neural networks for this task. While these methods
perform well, they don't provide satisfactory results. In this work, the idea of
metric-learning is extended with Support Vector Machines (SVM) working mechanism,
which is well known for generalization capabilities on a small dataset.
Furthermore, this paper presents an end-to-end learning framework for training
adaptive kernel SVMs, which eliminates the problem of choosing a correct kernel
and good features for SVMs. Next, the one-shot learning problem is redefined
for audio signals. Then the model was tested on vision task (using Omniglot
dataset) and speech task (using TIMIT dataset) as well. Actually, the algorithm
using Omniglot dataset improved accuracy from 98.1% to 98.5% on the one-shot
classification task and from 98.9% to 99.3% on the few-shot classification task.
****
What is image captioning made of?
Pranava Madhyastha,Josiah Wang,Lucia Specia
We hypothesize that end-to-end neural image captioning systems work seemingly well because they exploit and learn 'distributional similarity' in a multimodal feature space, by mapping a test image to similar training images in this space and generating a caption from the same space. To validate our hypothesis, we focus on the 'image' side of image captioning, and vary the input image representation but keep the RNN text generation model of a CNN-RNN constant. We propose a sparse bag-of-objects vector as an interpretable representation to investigate our distributional similarity hypothesis. We found that image captioning models (i) are capable of separating structure from noisy input representations; (ii) experience virtually no significant performance loss when a high dimensional representation is compressed to a lower dimensional space; (iii) cluster images with similar visual and linguistic information together; (iv) are heavily reliant on test sets with a similar distribution as the training set; (v) repeatedly generate the same captions by matching images and 'retrieving' a caption in the joint visual-textual space. Our experiments all point to one fact: that our distributional similarity hypothesis holds. We conclude that, regardless of the image representation, image captioning systems seem to match images and generate captions in a learned joint image-text semantic subspace.

****
The Context-Aware Learner
Conor Durkan,Amos Storkey,Harrison Edwards
One important aspect of generalization in machine learning involves reasoning about previously seen data in new settings. Such reasoning requires learning disentangled representations of data which are interpretable in isolation, but can also be combined in a new, unseen scenario. To this end, we introduce the context-aware learner, a model based on the variational autoencoding framework, which can learn such representations across data sets exhibiting a number of distinct contexts. Moreover, it is successfully able to combine these representations to generate data not seen at training time. The model enjoys an exponential increase in representational ability for a linear increase in context count. We demonstrate that the theory readily extends to a meta-learning setting such as this, and describe a fully unsupervised model in complete generality. Finally, we validate our approach using an adaptation with weak supervision.
****
Extending the Framework of Equilibrium Propagation to General Dynamics
Benjamin Scellier,Anirudh Goyal,Jonathan Binas,Thomas Mesnard,Yoshua Bengio

The biological plausibility of the backpropagation algorithm has long been doubted by neuroscientists. Two major reasons are that neurons would need to send two different types of signal in the forward and backward phases, and that pairs of neurons would need to communicate through symmetric bidirectional connections.
We present a simple two-phase learning procedure for fixed point recurrent networks that addresses both these issues.
In our model, neurons perform leaky integration and synaptic weights are updated through a local mechanism.
Our learning method extends the framework of Equilibrium Propagation to general dynamics, relaxing the requirement of an energy function.
As a consequence of this generalization, the algorithm does not compute the true gradient of the objective function,
but rather approximates it at a precision which is proven to be directly related to the degree of symmetry of the feedforward and feedback weights.
We show experimentally that the intrinsic properties of the system lead to alignment of the feedforward and feedback weights, and that our algorithm optimizes the objective function.
****
Towards Building Affect sensitive Word Distributions
Kushal Chawla,Sopan Khosla,Niyati Chhaya,Kokil Jaidka
Learning word representations from large available corpora relies on the distributional hypothesis that words present in similar contexts tend to have similar meanings. Recent work has shown that word representations learnt in this manner lack sentiment information which, fortunately, can be leveraged using external knowledge. Our work addresses the question: can affect lexica improve the word representations learnt from a corpus? In this work, we propose techniques to incorporate affect lexica, which capture fine-grained information about a word's psycholinguistic and emotional orientation, into the training process of Word2Vec SkipGram, Word2Vec CBOW and GloVe methods using a joint learning approach. We use affect scores from Warriner's affect lexicon to regularize the vector representations learnt from an unlabelled corpus. Our proposed method outperforms previously proposed methods on standard tasks for word similarity detection, outlier detection and sentiment detection. We also demonstrate the usefulness of our approach for a new task related to the prediction of formality, frustration and politeness in corporate communication.
****
Simple Nearest Neighbor Policy Method for Continuous Control Tasks
Elman Mansimov,Kyunghyun Cho
We design a new policy, called a nearest neighbor policy, that does not require any optimization for simple, low-dimensional continuous control tasks. As this policy does not require any optimization, it allows us to investigate the underlying difficulty of a task without being distracted by optimization difficulty of a learning algorithm. We propose two variants, one that retrieves an entire trajectory based on a pair of initial and goal states, and the other retrieving a partial trajectory based on a pair of current and goal states. We test the proposed policies on five widely-used benchmark continuous control tasks with a sparse reward: Reacher, Half Cheetah, Double Pendulum, Cart Pole and Mountain Car. We observe that the majority (the first four) of these tasks, which have been considered difficult, are easily solved by the proposed policies with high success rates, indicating that reported difficulties of them may have likely been due to the optimization difficulty. Our work suggests that it is necessary to evaluate any sophisticated policy learning algorithm on more challenging problems in order to truly assess the advances from them.
****
Unleashing the Potential of CNNs for Interpretable Few-Shot Learning
Boyang Deng,Qing Liu,Siyuan Qiao,Alan Yuille
Convolutional neural networks (CNNs) have been generally acknowledged as one of the driving forces for the advancement of computer vision. Despite their promising performances on many tasks, CNNs still face major obstacles on the road to achieving ideal machine intelligence. One is that CNNs are complex and hard to int

erpret. Another is that standard CNNs require large amounts of annotated data, w hich is sometimes very hard to obtain, and it is desirable to be able to learn t hem from few examples. In this work, we address these limitations of CNNs by dev eloping novel, simple, and interpretable models for few-shot learn- ing. Our mod els are based on the idea of encoding objects in terms of visual concepts, which are interpretable visual cues represented by the feature vectors within CNNs. W e first adapt the learning of visual concepts to the few-shot setting, and then uncover two key properties of feature encoding using visual concepts, which we c all category sensitivity and spatial pattern. Motivated by these properties, we present two intuitive models for the problem of few-shot learning. Experiments s how that our models achieve competitive performances, while being much more flex ible and interpretable than alternative state-of-the-art few-shot learning metho ds. We conclude that using visual concepts helps expose the natural capability o f CNNs for few-shot learning.
****
AirNet: a machine learning dataset for air quality forecasting
Songgang Zhao,Xingyuan Yuan,Da Xiao,Jianyuan Zhang,Zhouyuan Li
In the past decade, many urban areas in China have suffered from serious air pol lution problems, making air quality forecast a hot spot. Conventional approaches rely on numerical methods to estimate the pollutant concentration and require l ots of computing power. To solve this problem, we applied the widely used deep l earning methods. Deep learning requires large-scale datasets to train an effecti ve model. In this paper, we introduced a new dataset, entitled as AirNet, contai ning the 0.25 degree resolution grid map of mainland China, with more than two y ears of continued air quality measurement and meteorological data. We published this dataset as an open resource for machine learning researches and set up a ba seline of a 5-day air pollution forecast. The results of experiments demonstrate d that this dataset could facilitate the development of new algorithms on the ai r quality forecast.
****
Stochastic Hyperparameter Optimization through Hypernetworks
Jonathan Lorraine,David Duvenaud
Machine learning models are usually tuned by nesting optimization of model weigh ts inside the optimization of hyperparameters. We give a method to collapse thi s nested optimization into joint stochastic optimization of both weights and hyp erparameters. Our method trains a neural network to output approximately optima l weights as a function of hyperparameters. We show that our method converges t o locally optimal weights and hyperparameters for sufficiently large hypernets. We compare this method to standard hyperparameter optimization strategies and d emonstrate its effectiveness for tuning thousands of hyperparameters.
****
Orthogonal Recurrent Neural Networks with Scaled Cayley Transform
Kyle Helfrich,Devin Willmott,Qiang Ye
Recurrent Neural Networks (RNNs) are designed to handle sequential data but suff er from vanishing or exploding gradients. Recent work on Unitary Recurrent Neur al Networks (uRNNs) have been used to address this issue and in some cases, exce ed the capabilities of Long Short-Term Memory networks (LSTMs). We propose a si mpler and novel update scheme to maintain orthogonal recurrent weight matrices w ithout using complex valued matrices. This is done by parametrizing with a skew- symmetric matrix using the Cayley transform. Such a parametrization is unable to represent matrices with negative one eigenvalues, but this limitation is overco me by scaling the recurrent weight matrix by a diagonal matrix consisting of one s and negative ones. The proposed training scheme involves a straightforward gr adient calculation and update step. In several experiments, the proposed scaled Cayley orthogonal recurrent neural network (scoRNN) achieves superior results wi th fewer trainable parameters than other unitary RNNs.
****
Learning Sparse Neural Networks through L_0 Regularization
Christos Louizos,Max Welling,Diederik P. Kingma
We propose a practical method for L_0 norm regularization for neural networks:

pruning the network during training by encouraging weights to become exactly zero. Such regularization is interesting since (1) it can greatly speed up training and inference, and (2) it can improve generalization. AIC and BIC, well-known model selection criteria, are special cases of L_0 regularization. However, since the L_0 norm of weights is non-differentiable, we cannot incorporate it directly as a regularization term in the objective function. We propose a solution through the inclusion of a collection of non-negative stochastic gates, which collectively determine which weights to set to zero. We show that, somewhat surprisingly, for certain distributions over the gates, the expected L_0 regularized objective is differentiable with respect to the distribution parameters. We further propose the \emph{hard concrete} distribution for the gates, which is obtained by ``stretching'' a binary concrete distribution and then transforming its samples with a hard-sigmoid. The parameters of the distribution over the gates can then be jointly optimized with the original network parameters. As a result our method allows for straightforward and efficient learning of model structures with stochastic gradient descent and allows for conditional computation in a principled way. We perform various experiments to demonstrate the effectiveness of the resulting approach and regularizer.

****

Explicit Induction Bias for Transfer Learning with Convolutional Networks

Xuhong LI,Yves GRANDVALET,Franck DAVOINE

In inductive transfer learning, fine-tuning pre-trained convolutional networks substantially outperforms training from scratch.

When using fine-tuning, the underlying assumption is that the pre-trained model extracts generic features, which are at least partially relevant for solving the target task, but would be difficult to extract from the limited amount of data available on the target task.

However, besides the initialization with the pre-trained model and the early stopping, there is no mechanism in fine-tuning for retaining the features learned on the source task.

In this paper, we investigate several regularization schemes that explicitly promote the similarity of the final solution with the initial model.

We eventually recommend a simple L^2 penalty using the pre-trained model as a reference, and we show that this approach behaves much better than the standard scheme using weight decay on a partially frozen network.

****

Loss-aware Weight Quantization of Deep Networks

Lu Hou,James T. Kwok

The huge size of deep networks hinders their use in small computing devices. In this paper, we consider compressing the network by weight quantization. We extend a recently proposed loss-aware weight binarization scheme to ternarization, with possibly different scaling parameters for the positive and negative weights, and m-bit (where m > 2) quantization. Experiments on feedforward and recurrent neural networks show that the proposed scheme outperforms state-of-the-art weight quantization algorithms, and is as accurate (or even more accurate) than the full-precision network.

****

Clipping Free Attacks Against Neural Networks

Boussad ADDAD

During the last years, a remarkable breakthrough has been made in AI domain thanks to artificial deep neural networks that achieved a great success in many machine learning tasks in computer vision, natural language processing, speech recognition, malware detection and so on. However, they are highly vulnerable to easily crafted adversarial examples. Many investigations have pointed out this fact and different approaches have been proposed to generate attacks while adding a limited perturbation to the original data. The most robust known method so far is the so called C&W attack [1]. Nonetheless, a countermeasure known as feature squeezing coupled with ensemble defense showed that most of these attacks

can be destroyed [6]. In this paper, we present a new method we call Centered Initial Attack (CIA) whose advantage is twofold : first, it insures by construction the maximum perturbation to be smaller than a threshold fixed beforehand, without the clipping process that degrades the quality of attacks. Second, it is robust against recently introduced defenses such as feature squeezing, JPEG encoding and even against a voting ensemble of defenses. While its application is not limited to images, we illustrate this using five of the current best classif
iers
on ImageNet dataset among which two are adversarialy retrained on purpose to be robust against attacks. With a fixed maximum perturbation of only 1.5% on any pixel, around 80% of attacks (targeted) fool the voting ensemble defense and nearly 100% when the perturbation is only 6%. While this shows how it is difficu
lt
to defend against CIA attacks, the last section of the paper gives some guidelin
es
to limit their impact.
****

AUTOMATED DESIGN USING NEURAL NETWORKS AND GRADIENT DESCENT

Oliver Hennigh

We propose a novel method that makes use of deep neural networks and gradient de
cent to perform automated design on complex real world engineering tasks. Our ap
proach works by training a neural network to mimic the fitness function of a des
ign optimization task and then, using the differential nature of the neural netw
ork, perform gradient decent to maximize the fitness. We demonstrate this method
s effectiveness by designing an optimized heat sink and both 2D and 3D airfoils
that maximize the lift drag ratio under steady state flow conditions. We highlig
ht that our method has two distinct benefits over other automated design approac
hes. First, evaluating the neural networks prediction of fitness can be orders o
f magnitude faster then simulating the system of interest. Second, using gradien
t decent allows the design space to be searched much more efficiently then other
 gradient free methods. These two strengths work together to overcome some of th
e current shortcomings of automated design.
****

Novelty Detection with GAN

Mark Kliger,Shachar Fleishman

The ability of a classifier to recognize unknown inputs is important for many cl
assification-based systems. We discuss the problem of simultaneous classificatio
n and novelty detection, i.e. determining whether an input is from the known set
 of classes and from which specific class, or from an unknown domain and does no
t belong to any of the known classes. We propose a method based on the Generativ
e Adversarial Networks (GAN) framework. We show that a multi-class discriminator
 trained with a generator that generates samples from a mixture of nominal and n
ovel data distributions is the optimal novelty detector. We approximate that gen
erator with a mixture generator trained with the Feature Matching loss and empir
ically show that the proposed method outperforms conventional methods for novelt
y detection. Our findings demonstrate a simple, yet powerful new application of
the GAN framework for the task of novelty detection.
****

Don't Decay the Learning Rate, Increase the Batch Size

Samuel L. Smith,Pieter-Jan Kindermans,Chris Ying,Quoc V. Le

It is common practice to decay the learning rate. Here we show one can usually o
btain the same learning curve on both training and test sets by instead increasi
ng the batch size during training. This procedure is successful for stochastic g
radient descent (SGD), SGD with momentum, Nesterov momentum, and Adam. It reache
s equivalent test accuracies after the same number of training epochs, but with
fewer parameter updates, leading to greater parallelism and shorter training tim
es. We can further reduce the number of parameter updates by increasing the lear
ning rate ϵ and scaling the batch size $B \propto \epsilon$. Finally, o
ne can increase the momentum coefficient m and scale $B \propto 1/(1-m)$, alth
ough this tends to slightly reduce the test accuracy. Crucially, our techniques

allow us to repurpose existing training schedules for large batch training with no hyper-parameter tuning. We train ResNet-50 on ImageNet to 76.1% validation accuracy in under 30 minutes.
****

The (Un)reliability of saliency methods

Pieter-Jan Kindermans,Sara Hooker,Julius Adebayo,Kristof T. Schütt,Maximilian Alber,Sven Dähne,Dumitru Erhan,Been Kim

Saliency methods aim to explain the predictions of deep neural networks. These methods lack reliability when the explanation is sensitive to factors that do not contribute to the model prediction. We use a simple and common pre-processing step ---adding a mean shift to the input data--- to show that a transformation with no effect on the model can cause numerous methods to incorrectly attribute. We define input invariance as the requirement that a saliency method mirror the sensitivity of the model with respect to transformations of the input. We show, through several examples, that saliency methods that do not satisfy a input invariance property are unreliable and can lead to misleading and inaccurate attribution.
****

Neumann Optimizer: A Practical Optimization Algorithm for Deep Neural Networks

Shankar Krishnan,Ying Xiao,Rif. A. Saurous

Progress in deep learning is slowed by the days or weeks it takes to train large models. The natural solution of using more hardware is limited by diminishing returns, and leads to inefficient use of additional resources. In this paper, we present a large batch, stochastic optimization algorithm that is both faster than widely used algorithms for fixed amounts of computation, and also scales up substantially better as more computational resources become available. Our algorithm implicitly computes the inverse Hessian of each mini-batch to produce descent directions; we do so without either an explicit approximation to the Hessian or Hessian-vector products. We demonstrate the effectiveness of our algorithm by successfully training large ImageNet models (InceptionV3, ResnetV1-50, ResnetV1-101 and InceptionResnetV2) with mini-batch sizes of up to 32000 with no loss in validation error relative to current baselines, and no increase in the total number of steps. At smaller mini-batch sizes, our optimizer improves the validation error in these models by 0.8-0.9\%. Alternatively, we can trade off this accuracy to reduce the number of training steps needed by roughly 10-30\%. Our work is practical and easily usable by others -- only one hyperparameter (learning rate) needs tuning, and furthermore, the algorithm is as computationally cheap as the commonly used Adam optimizer.
****

Three factors influencing minima in SGD

Stanis■aw Jastrz■bski,Zac Kenton,Devansh Arpit,Nicolas Ballas,Asja Fischer,Amos Storkey,Yoshua Bengio

We study the statistical properties of the endpoint of stochastic gradient descent (SGD). We approximate SGD as a stochastic differential equation (SDE) and consider its Boltzmann Gibbs equilibrium distribution under the assumption of isotropic variance in loss gradients.. Through this analysis, we find that three factors – learning rate, batch size and the variance of the loss gradients – control the trade-off between the depth and width of the minima found by SGD, with wider minima favoured by a higher ratio of learning rate to batch size. In the equilibrium distribution only the ratio of learning rate to batch size appears, implying that it’s invariant under a simultaneous rescaling of each by the same amount.

We experimentally show how learning rate and batch size affect SGD from two perspectives: the endpoint of SGD and the dynamics that lead up to it. For the endpoint, the experiments suggest the endpoint of SGD is similar under simultaneous rescaling of batch size and learning rate, and also that a higher ratio leads to flatter minima, both findings are consistent with our theoretical analysis. We note experimentally that the dynamics also seem to be similar under the same rescaling of learning rate and batch size, which we explore showing that one can exchange batch size and learning rate in a cyclical learning rate schedule. Next, w

e illustrate how noise affects memorization, showing that high noise levels lead to better generalization. Finally, we find experimentally that the similarity under simultaneous rescaling of learning rate and batch size breaks down if the learning rate gets too large or the batch size gets too small.
****

Learning Approximate Inference Networks for Structured Prediction
Lifu Tu,Kevin Gimpel
Structured prediction energy networks (SPENs; Belanger & McCallum 2016) use neural network architectures to define energy functions that can capture arbitrary dependencies among parts of structured outputs. Prior work used gradient descent for inference, relaxing the structured output to a set of continuous variables and then optimizing the energy with respect to them. We replace this use of gradient descent with a neural network trained to approximate structured argmax inference. This
“inference network” outputs continuous values that we treat as the output structure. We develop large-margin training criteria for joint training of the structured energy function and inference network. On multi-label classification we report speed-ups
of 10-60x compared to (Belanger et al., 2017) while also improving accuracy. For sequence labeling with simple structured energies, our approach performs comparably to exact inference while being much faster at test time. We then demonstrate improved accuracy by augmenting the energy with a “label language model” that scores entire output label sequences, showing it can improve handling of long-distance dependencies in part-of-speech tagging. Finally, we show how inference networks can replace dynamic programming for test-time inference in conditional random fields, suggestive for their general use for fast inference in structured settings.
****

Beyond Finite Layer Neural Networks: Bridging Deep Architectures and Numerical Differential Equations
Yiping Lu,Aoxiao Zhong,Quanzheng Li,Bin Dong
Deep neural networks have become the state-of-the-art models in numerous machine learning tasks. However, general guidance to network architecture design is still missing. In our work, we bridge deep neural network design with numerical differential equations. We show that many effective networks, such as ResNet, PolyNet, FractalNet and RevNet, can be interpreted as different numerical discretizations of differential equations. This finding brings us a brand new perspective on the design of effective deep architectures. We can take advantage of the rich knowledge in numerical analysis to guide us in designing new and potentially more effective deep networks. As an example, we propose a linear multi-step architecture (LM-architecture) which is inspired by the linear multi-step method solving ordinary differential equations. The LM-architecture is an effective structure that can be used on any ResNet-like networks. In particular, we demonstrate that LM-ResNet and LM-ResNeXt (i.e. the networks obtained by applying the LM-architecture on ResNet and ResNeXt respectively) can achieve noticeably higher accuracy than ResNet and ResNeXt on both CIFAR and ImageNet with comparable numbers of trainable parameters. In particular, on both CIFAR and ImageNet, LM-ResNet/LM-ResNeXt can significantly compress (>50%) the original networks while maintaining a similar performance. This can be explained mathematically using the concept of modified equation from numerical analysis. Last but not least, we also establish a connection between stochastic control and noise injection in the training process which helps to improve generalization of the networks. Furthermore, by relating stochastic training strategy with stochastic dynamic system, we can easily apply stochastic training to the networks with the LM-architecture. As an example, we introduced stochastic depth to LM-ResNet and achieve significant improvement over the original LM-ResNet on CIFAR10.
****

Anytime Neural Network: a Versatile Trade-off Between Computation and Accuracy
Hanzhang Hu,Debadeepta Dey,Martial Hebert,J. Andrew Bagnell
We present an approach for anytime predictions in deep neural networks (DNNs). F

or each test sample, an anytime predictor produces a coarse result quickly, and then continues to refine it until the test-time computational budget is depleted. Such predictors can address the growing computational problem of DNNs by automatically adjusting to varying test-time budgets. In this work, we study a \emph{general} augmentation to feed-forward networks to form anytime neural networks (ANNs) via auxiliary predictions and losses. Specifically, we point out a blind-spot in recent studies in such ANNs: the importance of high final accuracy. In fact, we show on multiple recognition data-sets and architectures that by having near-optimal final predictions in small anytime models, we can effectively double the speed of large ones to reach corresponding accuracy level. We achieve such speed-up with simple weighting of anytime losses that oscillate during training. We also assemble a sequence of exponentially deepening ANNs, to achieve both theoretically and practically near-optimal anytime results at any budget, at the cost of a constant fraction of additional consumed budget.
****
Stabilizing Gradients for Deep Neural Networks via Efficient SVD Parameterization
Jiong Zhang,Qi Lei,Inderjit S. Dhillon
Vanishing and exploding gradients are two of the main obstacles in training deep neural networks, especially in capturing long range dependencies in recurrent neural networks (RNNs). In this paper, we present an efficient parametrization of the transition matrix of an RNN that allows us to stabilize the gradients that arise in its training. Specifically, we parameterize the transition matrix by its singular value decomposition (SVD), which allows us to explicitly track and control its singular values. We attain efficiency by using tools that are common in numerical linear algebra, namely Householder reflectors for representing the orthogonal matrices that arise in the SVD. By explicitly controlling the singular values, our proposed svdRNN method allows us to easily solve the exploding gradient problem and we observe that it empirically solves the vanishing gradient issue to a large extent. We note that the SVD parameterization can be used for any rectangular weight matrix, hence it can be easily extended to any deep neural network, such as a multi-layer perceptron. Theoretically, we demonstrate that our parameterization does not lose any expressive power, and show how it potentially makes the optimization process easier. Our extensive experimental results also demonstrate that the proposed framework converges faster, and has good generalization, especially when the depth is large.

****
IVE-GAN: Invariant Encoding Generative Adversarial Networks
Robin Winter,Djork-Arnè Clevert
Generative adversarial networks (GANs) are a powerful framework for generative tasks. However, they are difficult to train and tend to miss modes of the true data generation process. Although GANs can learn a rich representation of the covered modes of the data in their latent space, the framework misses an inverse mapping from data to this latent space. We propose Invariant Encoding Generative Adversarial Networks (IVE-GANs), a novel GAN framework that introduces such a mapping for individual samples from the data by utilizing features in the data which are invariant to certain transformations. Since the model maps individual samples to the latent space, it naturally encourages the generator to cover all modes. We demonstrate the effectiveness of our approach in terms of generative performance and learning rich representations on several datasets including common benchmark image generation tasks.
****
When is a Convolutional Filter Easy to Learn?
Simon S. Du,Jason D. Lee,Yuandong Tian
We analyze the convergence of (stochastic) gradient descent algorithm for learning a convolutional filter with Rectified Linear Unit (ReLU) activation function. Our analysis does not rely on any specific form of the input distribution and our proofs only use the definition of ReLU, in contrast with previous works that are restricted to standard Gaussian input. We show that (stochastic) gradient de

scent with random initialization can learn the convolutional filter in polynomia
l time and the convergence rate depends on the smoothness of the input distribut
ion and the closeness of patches. To the best of our knowledge, this is the firs
t recovery guarantee of gradient-based algorithms for convolutional filter on no
n-Gaussian input distributions. Our theory also justifies the two-stage learning
 rate strategy in deep neural networks. While our focus is theoretical, we also
present experiments that justify our theoretical findings.
****

An information-theoretic analysis of deep latent-variable models
Alex Alemi,Ben Poole,Ian Fischer,Josh Dillon,Rif A. Saurus,Kevin Murphy
We present an information-theoretic framework for understanding trade-offs in un
supervised learning of deep latent-variables models using variational inference.
 This framework emphasizes the need to consider latent-variable models along two
 dimensions: the ability to reconstruct inputs (distortion) and the communicatio
n cost (rate). We derive the optimal frontier of generative models in the two-di
mensional rate-distortion plane, and show how the standard evidence lower bound
objective is insufficient to select between points along this frontier. However,
 by performing targeted optimization to learn generative models with different r
ates, we are able to learn many models that can achieve similar generative perfo
rmance but make vastly different trade-offs in terms of the usage of the latent
variable. Through experiments on MNIST and Omniglot with a variety of architectu
res, we show how our framework sheds light on many recent proposed extensions to
 the variational autoencoder family.
****

Fast and Accurate Text Classification: Skimming, Rereading and Early Stopping
Keyi Yu,Yang Liu,Alexander G. Schwing,Jian Peng
Recent advances in recurrent neural nets (RNNs) have shown much promise in many
applications in natural language processing. For most of these tasks, such as se
ntiment analysis of customer reviews, a recurrent neural net model parses the en
tire review before forming a decision. We argue that reading the entire input is
 not always necessary in practice, since a lot of reviews are often easy to clas
sify, i.e., a decision can be formed after reading some crucial sentences or wor
ds in the provided text. In this paper, we present an approach of fast reading f
or text classification. Inspired by several well-known human reading techniques,
 our approach implements an intelligent recurrent agent which evaluates the impo
rtance of the current snippet in order to decide whether to make a prediction, o
r to skip some texts, or to re-read part of the sentence. Our agent uses an RNN
module to encode information from the past and the current tokens, and applies a
 policy module to form decisions. With an end-to-end training algorithm based on
 policy gradient, we train and test our agent on several text classification dat
asets and achieve both higher efficiency and better accuracy compared to previou
s approaches.

****

Learning Dynamic State Abstractions for Model-Based Reinforcement Learning
Lars Buesing,Theophane Weber,Sebastien Racaniere,S. M. Ali Eslami,Danilo Rezende
,David Reichert,Fabio Viola,Frederic Besse,Karol Gregor,Demis Hassabis,Daan Wier
stra
A key challenge in model-based reinforcement learning (RL) is to synthesize comp
utationally efficient and accurate environment models. We show that carefully de
signed models that learn predictive and compact state representations, also call
ed state-space models, substantially reduce the computational costs for predicti
ng outcomes of sequences of actions. Extensive experiments establish that state-
space models accurately capture the dynamics of Atari games from the Arcade Lear
ning Environment (ALE) from raw pixels. Furthermore, RL agents that use Monte-Ca
rlo rollouts of these models as features for decision making outperform strong m
odel-free baselines on the game MS_PACMAN, demonstrating the benefits of plannin
g using learned dynamic state abstractions.
****

Unsupervised Learning of Entailment-Vector Word Embeddings

James Henderson
Entailment vectors are a principled way to encode in a vector what information is known and what is unknown. They are designed to model relations where one vector should include all the information in another vector, called entailment. This paper investigates the unsupervised learning of entailment vectors for the semantics of words. Using simple entailment-based models of the semantics of words in text (distributional semantics), we induce entailment-vector word embeddings which outperform the best previous results for predicting entailment between words, in unsupervised and semi-supervised experiments on hyponymy.

****
Character Level Based Detection of DGA Domain Names
Bin Yu,Jie Pan,Jiaming Hu,Anderson Nascimento,Martine De Cock
Recently several different deep learning architectures have been proposed that take a string of characters as the raw input signal and automatically derive features for text classification. Little studies are available that compare the effectiveness of these approaches for character based text classification with each other. In this paper we perform such an empirical comparison for the important cybersecurity problem of DGA detection: classifying domain names as either benign vs. produced by malware (i.e., by a Domain Generation Algorithm). Training and evaluating on a dataset with 2M domain names shows that there is surprisingly little difference between various convolutional neural network (CNN) and recurrent neural network (RNN) based architectures in terms of accuracy, prompting a preference for the simpler architectures, since they are faster to train and less prone to overfitting.
****
Exploring Asymmetric Encoder-Decoder Structure for Context-based Sentence Representation Learning
Shuai Tang,Hailin Jin,Chen Fang,Zhaowen Wang,Virginia R. de Sa
Context information plays an important role in human language understanding, and it is also useful for machines to learn vector representations of language. In this paper, we explore an asymmetric encoder-decoder structure for unsupervised context-based sentence representation learning. As a result, we build an encoder-decoder architecture with an RNN encoder and a CNN decoder, and we show that neither an autoregressive decoder nor an RNN decoder is required. We further combine a suite of effective designs to significantly improve model efficiency while also achieving better performance. Our model is trained on two different large unlabeled corpora, and in both cases transferability is evaluated on a set of downstream language understanding tasks. We empirically show that our model is simple and fast while producing rich sentence representations that excel in downstream tasks.
****
Gaussian Process Neurons
Sebastian Urban,Patrick van der Smagt
We propose a method to learn stochastic activation functions for use in probabilistic neural networks.
First, we develop a framework to embed stochastic activation functions based on Gaussian processes in probabilistic neural networks.
Second, we analytically derive expressions for the propagation of means and covariances in such a network, thus allowing for an efficient implementation and training without the need for sampling.
Third, we show how to apply variational Bayesian inference to regularize and efficiently train this model.
The resulting model can deal with uncertain inputs and implicitly provides an estimate of the confidence of its predictions.
Like a conventional neural network it can scale to datasets of arbitrary size and be extended with convolutional and recurrent connections, if desired.
****
Finding ReMO (Related Memory Object): A Simple neural architecture for Text based Reasoning

Jihyung Moon,Hyochang Yang,Sungzoon Cho
Memory Network based models have shown a remarkable progress on the task of rela tional reasoning.
Recently, a simpler yet powerful neural network module called Relation Network (RN) has been introduced.
Despite its architectural simplicity, the time complexity of relation network gr ows quadratically with data, hence limiting its application to tasks with a larg e-scaled memory.
We introduce Related Memory Network, an end-to-end neural network architecture e xploiting both memory network and relation network structures.
We follow memory network's four components while each component operates similar to the relation network without taking a pair of objects.
As a result, our model is as simple as RN but the computational complexity is re duced to linear time.
It achieves the state-of-the-art results in jointly trained bAbI-10k story-based question answering and bAbI dialog dataset.
****
Tree-to-tree Neural Networks for Program Translation
Xinyun Chen,Chang Liu,Dawn Song
Program translation is an important tool to migrate legacy code in one language into an ecosystem built in a different language. In this work, we are the first to consider employing deep neural networks toward tackling this problem. We obse rve that program translation is a modular procedure, in which a sub-tree of the source tree is translated into the corresponding target sub-tree at each step. T o capture this intuition, we design a tree-to-tree neural network as an encoder-decoder architecture to translate a source tree into a target one. Meanwhile, we develop an attention mechanism for the tree-to-tree model, so that when the dec oder expands one non-terminal in the target tree, the attention mechanism locate s the corresponding sub-tree in the source tree to guide the expansion of the de coder. We evaluate the program translation capability of our tree-to-tree model against several state-of-the-art approaches. Compared against other neural trans lation models, we observe that our approach is consistently better than the base lines with a margin of up to 15 points. Further, our approach can improve the pr evious state-of-the-art program translation approaches by a margin of 20 points on the translation of real-world projects.
****
Can Deep Reinforcement Learning solve Erdos-Selfridge-Spencer Games?
Maithra Raghu,Alex Irpan,Jacob Andreas,Robert Kleinberg,Quoc Le,Jon Kleinberg
Deep reinforcement learning has achieved many recent successes, but our understa nding of its strengths and limitations is hampered by the lack of rich environme nts in which we can fully characterize optimal behavior, and correspondingly dia gnose individual actions against such a characterization.

Here we consider a family of combinatorial games, arising from work of Erdos, Se lfridge, and Spencer, and we propose their use as environments for evaluating an d comparing different approaches to reinforcement learning. These games have a n umber of appealing features: they are challenging for current learning approache s, but they form (i) a low-dimensional, simply parametrized environment where (i i) there is a linear closed form solution for optimal behavior from any state, a nd (iii) the difficulty of the game can be tuned by changing environment paramet ers in an interpretable way. We use these Erdos-Selfridge-Spencer games not only to compare different algorithms, but also to compare approaches based on superv ised and reinforcement learning, to analyze the power of multi-agent approaches in improving performance, and to evaluate generalization to environments outside the training set.
****
Variance-based Gradient Compression for Efficient Distributed Deep Learning
Yusuke Tsuzuku,Hiroto Imachi,Takuya Akiba
Due to the substantial computational cost, training state-of-the-art deep neural networks for large-scale datasets often requires distributed training using mul

tiple computation workers. However, by nature, workers need to frequently commun icate gradients, causing severe bottlenecks, especially on lower bandwidth conne ctions. A few methods have been proposed to compress gradient for efficient comm unication, but they either suffer a low compression ratio or significantly harm the resulting model accuracy, particularly when applied to convolutional neural networks. To address these issues, we propose a method to reduce the communicati on overhead of distributed deep learning. Our key observation is that gradient u pdates can be delayed until an unambiguous (high amplitude, low variance) gradie nt has been calculated. We also present an efficient algorithm to compute the va riance and prove that it can be obtained with negligible additional cost. We exp erimentally show that our method can achieve very high compression ratio while m aintaining the result model accuracy. We also analyze the efficiency using compu tation and communication cost models and provide the evidence that this method e nables distributed deep learning for many scenarios with commodity environments.
****

Self-ensembling for visual domain adaptation
Geoff French,Michal Mackiewicz,Mark Fisher
This paper explores the use of self-ensembling for visual domain adaptation prob lems. Our technique is derived from the mean teacher variant (Tarvainen et. al 2 017) of temporal ensembling (Laine et al. 2017), a technique that achieved state of the art results in the area of semi-supervised learning. We introduce a numb er of modifications to their approach for challenging domain adaptation scenario s and evaluate its effectiveness. Our approach achieves state of the art results in a variety of benchmarks, including our winning entry in the VISDA-2017 visua l domain adaptation challenge. In small image benchmarks, our algorithm not only outperforms prior art, but can also achieve accuracy that is close to that of a classifier trained in a supervised fashion.
****

Boosting Dilated Convolutional Networks with Mixed Tensor Decompositions
Nadav Cohen,Ronen Tamari,Amnon Shashua
The driving force behind deep networks is their ability to compactly represent r ich classes of functions. The primary notion for formally reasoning about this p henomenon is expressive efficiency, which refers to a situation where one networ k must grow unfeasibly large in order to replicate functions of another. To date , expressive efficiency analyses focused on the architectural feature of depth, showing that deep networks are representationally superior to shallow ones. In t his paper we study the expressive efficiency brought forth by connectivity, moti vated by the observation that modern networks interconnect their layers in elabo rate ways. We focus on dilated convolutional networks, a family of deep models d elivering state of the art performance in sequence processing tasks. By introduc ing and analyzing the concept of mixed tensor decompositions, we prove that inte rconnecting dilated convolutional networks can lead to expressive efficiency. In particular, we show that even a single connection between intermediate layers c an already lead to an almost quadratic gap, which in large-scale settings typica lly makes the difference between a model that is practical and one that is not. Empirical evaluation demonstrates how the expressive efficiency of connectivity, similarly to that of depth, translates into gains in accuracy. This leads us to believe that expressive efficiency may serve a key role in developing new tools for deep network design.
****

Automatically Inferring Data Quality for Spatiotemporal Forecasting
Sungyong Seo,Arash Mohegh,George Ban-Weiss,Yan Liu
Spatiotemporal forecasting has become an increasingly important prediction task in machine learning and statistics due to its vast applications, such as climate modeling, traffic prediction, video caching predictions, and so on. While numer ous studies have been conducted, most existing works assume that the data from d ifferent sources or across different locations are equally reliable. Due to cost , accessibility, or other factors, it is inevitable that the data quality could vary, which introduces significant biases into the model and leads to unreliable prediction results. The problem could be exacerbated in black-box prediction mo

dels, such as deep neural networks. In this paper, we propose a novel solution that can automatically infer data quality levels of different sources through local variations of spatiotemporal signals without explicit labels. Furthermore, we integrate the estimate of data quality level with graph convolutional networks to exploit their efficient structures. We evaluate our proposed method on forecasting temperatures in Los Angeles.
****

Learning To Generate Reviews and Discovering Sentiment
Alec Radford,Rafal Jozefowicz,Ilya Sutskever
We explore the properties of byte-level recurrent language models. When given sufficient amounts of capacity, training data, and compute time, the representations learned by these models include disentangled features corresponding to high-level concepts. Specifically, we find a single unit which performs sentiment analysis. These representations, learned in an unsupervised manner, achieve state of the art on the binary subset of the Stanford Sentiment Treebank. They are also very data efficient. When using only a handful of labeled examples, our approach matches the performance of strong baselines trained on full datasets. We also demonstrate the sentiment unit has a direct influence on the generative process of the model. Simply fixing its value to be positive or negative generates samples with the corresponding positive or negative sentiment.
****

The Principle of Logit Separation
Gil Keren,Sivan Sabato,Björn Schuller
We consider neural network training, in applications in which there are many possible classes, but at test-time, the task is to identify only whether the given example belongs to a specific class, which can be different in different applications of the classifier. For instance, this is the case in an image search engine. We consider the Single Logit Classification (SLC) task: training the network so that at test-time, it would be possible to accurately identify if the example belongs to a given class, based only on the output logit for this class.
We propose a natural principle, the Principle of Logit Separation, as a guideline for choosing and designing losses suitable for the SLC.
We show that the cross-entropy loss function is not aligned with the Principle of Logit Separation. In contrast, there are known loss functions, as well as novel batch loss functions that we propose, which are aligned with this principle. In total, we study seven loss functions.
Our experiments show that indeed in almost all cases, losses that are aligned with Principle of Logit Separation obtain a 20%-35% relative performance improvement in the SLC task, compared to losses that are not aligned with it. We therefore conclude that the Principle of Logit Separation sheds light on an important property of the most common loss functions used by neural network classifiers.
****

Still not systematic after all these years: On the compositional skills of sequence-to-sequence recurrent networks
Brenden Lake,Marco Baroni
Humans can understand and produce new utterances effortlessly, thanks to their systematic compositional skills. Once a person learns the meaning of a new verb "dax," he or she can immediately understand the meaning of "dax twice" or "sing and dax." In this paper, we introduce the SCAN domain, consisting of a set of simple compositional navigation commands paired with the corresponding action sequences. We then test the zero-shot generalization capabilities of a variety of recurrent neural networks (RNNs) trained on SCAN with sequence-to-sequence methods. We find that RNNs can generalize well when the differences between training and test commands are small, so that they can apply "mix-and-match" strategies to solve the task. However, when generalization requires systematic compositional skills (as in the "dax" example above), RNNs fail spectacularly. We conclude with a proof-of-concept experiment in neural machine translation, supporting the conjecture that lack of systematicity is an important factor explaining why neural networks need very large training sets.
****

Learning Less-Overlapping Representations
Hongbao Zhang,Pengtao Xie,Eric Xing
In representation learning (RL), how to make the learned representations easy to interpret and less overfitted to training data are two important but challenging issues. To address these problems, we study a new type of regularization approach that encourages the supports of weight vectors in RL models to have small overlap, by simultaneously promoting near-orthogonality among vectors and sparsity of each vector. We apply the proposed regularizer to two models: neural networks (NNs) and sparse coding (SC), and develop an efficient ADMM-based algorithm for regularized SC. Experiments on various datasets demonstrate that weight vectors learned under our regularizer are more interpretable and have better generalization performance.
****
Learning Differentially Private Recurrent Language Models
H. Brendan McMahan,Daniel Ramage,Kunal Talwar,Li Zhang
We demonstrate that it is possible to train large recurrent language models with user-level differential privacy guarantees with only a negligible cost in predictive accuracy. Our work builds on recent advances in the training of deep networks on user-partitioned data and privacy accounting for stochastic gradient descent. In particular, we add user-level privacy protection to the federated averaging algorithm, which makes large step updates from user-level data. Our work demonstrates that given a dataset with a sufficiently large number of users (a requirement easily met by even small internet-scale datasets), achieving differential privacy comes at the cost of increased computation, rather than in decreased utility as in most prior work. We find that our private LSTM language models are quantitatively and qualitatively similar to un-noised models when trained on a large dataset.
****
Style Memory: Making a Classifier Network Generative
Rey Wiyatno,Jeff Orchard
Deep networks have shown great performance in classification tasks. However, the parameters learned by the classifier networks usually discard stylistic information of the input, in favour of information strictly relevant to classification. We introduce a network that has the capacity to do both classification and reconstruction by adding a "style memory" to the output layer of the network. We also show how to train such a neural network as a deep multi-layer autoencoder, jointly minimizing both classification and reconstruction losses. The generative capacity of our network demonstrates that the combination of style-memory neurons with the classifier neurons yield good reconstructions of the inputs when the classification is correct. We further investigate the nature of the style memory, and how it relates to composing digits and letters.
****
Bayesian Uncertainty Estimation for Batch Normalized Deep Networks
Mattias Teye,Hossein Azizpour,Kevin Smith
Deep neural networks have led to a series of breakthroughs, dramatically improving the state-of-the-art in many domains. The techniques driving these advances, however, lack a formal method to account for model uncertainty. While the Bayesian approach to learning provides a solid theoretical framework to handle uncertainty, inference in Bayesian-inspired deep neural networks is difficult. In this paper, we provide a practical approach to Bayesian learning that relies on a regularization technique found in nearly every modern network, batch normalization. We show that training a deep network using batch normalization is equivalent to approximate inference in Bayesian models, and we demonstrate how this finding allows us to make useful estimates of the model uncertainty. Using our approach, it is possible to make meaningful uncertainty estimates using conventional architectures without modifying the network or the training procedure. Our approach is thoroughly validated in a series of empirical experiments on different tasks and using various measures, showing it to outperform baselines on a majority of datasets with strong statistical significance.
****

Auto-Encoding Sequential Monte Carlo
Tuan Anh Le,Maximilian Igl,Tom Rainforth,Tom Jin,Frank Wood
We build on auto-encoding sequential Monte Carlo (AESMC): a method for model and proposal learning based on maximizing the lower bound to the log marginal likelihood in a broad family of structured probabilistic models. Our approach relies on the efficiency of sequential Monte Carlo (SMC) for performing inference in structured probabilistic models and the flexibility of deep neural networks to model complex conditional probability distributions. We develop additional theoretical insights and introduce a new training procedure which improves both model and proposal learning. We demonstrate that our approach provides a fast, easy-to-implement and scalable means for simultaneous model learning and proposal adaptation in deep generative models.
****
Variance Regularizing Adversarial Learning
Karan Grewal,R Devon Hjelm,Yoshua Bengio
We study how, in generative adversarial networks, variance in the discriminator's output affects the generator's ability to learn the data distribution. In particular, we contrast the results from various well-known techniques for training GANs when the discriminator is near-optimal and updated multiple times per update to the generator. As an alternative, we propose an additional method to train GANs by explicitly modeling the discriminator's output as a bi-modal Gaussian distribution over the real/fake indicator variables. In order to do this, we train the Gaussian classifier to match the target bi-modal distribution implicitly through meta-adversarial training. We observe that our new method, when trained together with a strong discriminator, provides meaningful, non-vanishing gradients.
****
Dynamic Integration of Background Knowledge in Neural NLU Systems
Dirk Weissenborn,Tomas Kocisky,Chris Dyer
Common-sense or background knowledge is required to understand natural language, but in most neural natural language understanding (NLU) systems, the requisite background knowledge is indirectly acquired from static corpora. We develop a new reading architecture for the dynamic integration of explicit background knowledge in NLU models. A new task-agnostic reading module provides refined word representations to a task-specific NLU architecture by processing background knowledge in the form of free-text statements, together with the task-specific inputs. Strong performance on the tasks of document question answering (DQA) and recognizing textual entailment (RTE) demonstrate the effectiveness and flexibility of our approach. Analysis shows that our models learn to exploit knowledge selectively and in a semantically appropriate way.
****
TRUNCATED HORIZON POLICY SEARCH: COMBINING REINFORCEMENT LEARNING & IMITATION LEARNING
Wen Sun,J. Andrew Bagnell,Byron Boots
In this paper, we propose to combine imitation and reinforcement learning via the idea of reward shaping using an oracle. We study the effectiveness of the near- optimal cost-to-go oracle on the planning horizon and demonstrate that the cost- to-go oracle shortens the learner’s planning horizon as function of its accuracy: a globally optimal oracle can shorten the planning horizon to one, leading to a one- step greedy Markov Decision Process which is much easier to optimize, while an oracle that is far away from the optimality requires planning over a longer horizon to achieve near-optimal performance. Hence our new insight bridges the gap and interpolates between imitation learning and reinforcement learning. Motivated by the above mentioned insights, we propose Truncated HORizon Policy Search (THOR), a method that focuses on searching for policies that maximize the total reshaped reward over a finite planning horizon when the oracle is sub-optimal. We experimentally demonstrate that a gradient-based implementation of THOR can achieve superior performance compared to RL baselines and IL baselines even when the oracle is sub-optimal.
****

Emergence of Linguistic Communication from Referential Games with Symbolic and Pixel Input
Angeliki Lazaridou,Karl Moritz Hermann,Karl Tuyls,Stephen Clark
The ability of algorithms to evolve or learn (compositional) communication protocols has traditionally been studied in the language evolution literature through the use of emergent communication tasks. Here we scale up this research by using contemporary deep learning methods and by training reinforcement-learning neural network agents on referential communication games. We extend previous work, in which agents were trained in symbolic environments, by developing agents which are able to learn from raw pixel data, a more challenging and realistic input representation. We find that the degree of structure found in the input data affects the nature of the emerged protocols, and thereby corroborate the hypothesis that structured compositional language is most likely to emerge when agents perceive the world as being structured.
****
YellowFin and the Art of Momentum Tuning
Jian Zhang,Ioannis Mitliagkas,Christopher Re
Hyperparameter tuning is one of the most time-consuming workloads in deep learning. State-of-the-art optimizers, such as AdaGrad, RMSProp and Adam, reduce this labor by adaptively tuning an individual learning rate for each variable. Recently researchers have shown renewed interest in simpler methods like momentum SGD as they may yield better results. Motivated by this trend, we ask: can simple adaptive methods, based on SGD perform as well or better? We revisit the momentum SGD algorithm and show that hand-tuning a single learning rate and momentum makes it competitive with Adam. We then analyze its robustness to learning rate misspecification and objective curvature variation. Based on these insights, we design YellowFin, an automatic tuner for momentum and learning rate in SGD. YellowFin optionally uses a negative-feedback loop to compensate for the momentum dynamics in asynchronous settings on the fly. We empirically show YellowFin can converge in fewer iterations than Adam on ResNets and LSTMs for image recognition, language modeling and constituency parsing, with a speedup of up to 3.28x in synchronous and up to 2.69x in asynchronous settings.
****
Adversarially Regularized Autoencoders
Junbo (Jake) Zhao,Yoon Kim,Kelly Zhang,Alexander M. Rush,Yann LeCun
While autoencoders are a key technique in representation learning for continuous structures, such as images or wave forms, developing general-purpose autoencoders for discrete structures, such as text sequence or discretized images, has proven to be more challenging. In particular, discrete inputs make it more difficult to learn a smooth encoder that preserves the complex local relationships in the input space. In this work, we propose an adversarially regularized autoencoder (ARAE) with the goal of learning more robust discrete-space representations. ARAE jointly trains both a rich discrete-space encoder, such as an RNN, and a simpler continuous space generator function, while using generative adversarial network (GAN) training to constrain the distributions to be similar. This method yields a smoother contracted code space that maps similar inputs to nearby codes, and also an implicit latent variable GAN model for generation. Experiments on text and discretized images demonstrate that the GAN model produces clean interpolations and captures the multimodality of the original space, and that the autoencoder produces improvements in semi-supervised learning as well as state-of-the-art results in unaligned text style transfer task using only a shared continuous-space representation.
****
Network Signatures from Image Representation of Adjacency Matrices: Deep/Transfer Learning for Subgraph Classification
Kshiteesh Hegde,Malik Magdon-Ismail,Ram Ramanathan,Bishal Thapa
We propose a novel subgraph image representation for classification of network fragments with the target being their parent networks. The graph image representation is based on 2D image embeddings of adjacency matrices. We use this image representation in two modes. First, as the input to a machine learning algorithm.

Second, as the input to a pure transfer learner. Our conclusions from multiple d
atasets are that
1. deep learning using structured image features performs the best compared to g
raph kernel and classical features based methods; and,
2. pure transfer learning works effectively with minimum interference from the u
ser and is robust against small data.

****
On the insufficiency of existing momentum schemes for Stochastic Optimization
Rahul Kidambi,Praneeth Netrapalli,Prateek Jain,Sham M. Kakade
Momentum based stochastic gradient methods such as heavy ball (HB) and Nesterov'
s accelerated gradient descent (NAG) method are widely used in practice for trai
ning deep networks and other supervised learning models, as they often provide s
ignificant improvements over stochastic gradient descent (SGD). Rigorously speak
ing, fast gradient methods have provable improvements over gradient descent only
 for the deterministic case, where the gradients are exact. In the stochastic ca
se, the popular explanations for their wide applicability is that when these fas
t gradient methods are applied in the stochastic case, they partially mimic thei
r exact gradient counterparts, resulting in some practical gain. This work provi
des a counterpoint to this belief by proving that there exist simple problem ins
tances where these methods cannot outperform SGD despite the best setting of its
 parameters. These negative problem instances are, in an informal sense, generic
; they do not look like carefully constructed pathological instances. These resu
lts suggest (along with empirical evidence) that HB or NAG's practical performan
ce gains are a by-product of minibatching.

Furthermore, this work provides a viable (and provable) alternative, which, on t
he same set of problem instances, significantly improves over HB, NAG, and SGD's
 performance. This algorithm, referred to as Accelerated Stochastic Gradient Des
cent (ASGD), is a simple to implement stochastic algorithm, based on a relativel
y less popular variant of Nesterov's Acceleration. Extensive empirical results i
n this paper show that ASGD has performance gains over HB, NAG, and SGD. The cod
e for implementing the ASGD Algorithm can be found at https://github.com/rahulki
dambi/AccSGD.

****
Ask the Right Questions: Active Question Reformulation with Reinforcement Learni
ng
Christian Buck,Jannis Bulian,Massimiliano Ciaramita,Wojciech Gajewski,Andrea Ges
mundo,Neil Houlsby,Wei Wang.
We frame Question Answering (QA) as a Reinforcement Learning task, an approach t
hat we call Active Question Answering.

We propose an agent that sits between the user and a black box QA system and lea
rns to reformulate questions to elicit the best possible answers. The agent prob
es the system with, potentially many, natural language reformulations of an init
ial question and aggregates the returned evidence to yield the best answer.

The reformulation system is trained end-to-end to maximize answer quality using
policy gradient. We evaluate on SearchQA, a dataset of complex questions extract
ed from Jeopardy!. The agent outperforms a state-of-the-art base model, playing
the role of the environment, and other benchmarks.

We also analyze the language that the agent has learned while interacting with t
he question answering system. We find that successful question reformulations lo
ok quite different from natural language paraphrases. The agent is able to disco
ver non-trivial reformulation strategies that resemble classic information retri
eval techniques such as term re-weighting (tf-idf) and stemming.
****
LSD-Net: Look, Step and Detect for Joint Navigation and Multi-View Recognition w

ith Deep Reinforcement Learning
N dinesh reddy
Multi-view recognition is the task of classifying an object from multi-view image sequences. Instead of using a single-view for classification, humans generally navigate around a target object to learn its multi-view representation. Motivated by this human behavior, the next best view can be learned by combining object recognition with navigation in complex environments. Since deep reinforcement learning has proven successful in navigation tasks, we propose a novel multi-task reinforcement learning framework for joint multi-view recognition and navigation. Our method uses a hierarchical action space for multi-task reinforcement learning. The framework was evaluated with an environment created from the ModelNet40 dataset. Our results show improvements on object recognition and demonstrate human-like behavior on navigation.
****
Recurrent Auto-Encoder Model for Multidimensional Time Series Representation
Timothy Wong,Zhiyuan Luo
Recurrent auto-encoder model can summarise sequential data through an encoder structure into a fixed-length vector and then reconstruct into its original sequential form through the decoder structure. The summarised information can be used to represent time series features. In this paper, we propose relaxing the dimensionality of the decoder output so that it performs partial reconstruction. The fixed-length vector can therefore represent features only in the selected dimensions. In addition, we propose using rolling fixed window approach to generate samples. The change of time series features over time can be summarised as a smooth trajectory path. The fixed-length vectors are further analysed through additional visualisation and unsupervised clustering techniques.

This proposed method can be applied in large-scale industrial processes for sensors signal analysis purpose where clusters of the vector representations can be used to reflect the operating states of selected aspects of the industrial system.
****
Sample-Efficient Deep Reinforcement Learning via Episodic Backward Update
Su Young Lee,Sungik Choi,Sae-Young Chung
We propose Episodic Backward Update - a new algorithm to boost the performance of a deep reinforcement learning agent by fast reward propagation. In contrast to the conventional use of the replay memory with uniform random sampling, our agent samples a whole episode and successively propagates the value of a state into its previous states. Our computationally efficient recursive algorithm allows sparse and delayed rewards to propagate effectively throughout the sampled episode. We evaluate our algorithm on 2D MNIST Maze Environment and 49 games of the Atari 2600 Environment and show that our agent improves sample efficiency with a competitive computational cost.
****
Understanding image motion with group representations
Andrew Jaegle,Stephen Phillips,Daphne Ippolito,Kostas Daniilidis
Motion is an important signal for agents in dynamic environments, but learning to represent motion from unlabeled video is a difficult and underconstrained problem. We propose a model of motion based on elementary group properties of transformations and use it to train a representation of image motion. While most methods of estimating motion are based on pixel-level constraints, we use these group properties to constrain the abstract representation of motion itself. We demonstrate that a deep neural network trained using this method captures motion in both synthetic 2D sequences and real-world sequences of vehicle motion, without requiring any labels. Networks trained to respect these constraints implicitly identify the image characteristic of motion in different sequence types. In the context of vehicle motion, this method extracts information useful for localization, tracking, and odometry. Our results demonstrate that this representation is useful for learning motion in the general setting where explicit labels are difficult to obtain.

****
Learning Latent Permutations with Gumbel-Sinkhorn Networks
Gonzalo Mena,David Belanger,Scott Linderman,Jasper Snoek
Permutations and matchings are core building blocks in a variety of latent varia ble models, as they allow us to align, canonicalize, and sort data. Learning in such models is difficult, however, because exact marginalization over these comb inatorial objects is intractable. In response, this paper introduces a collectio n of new methods for end-to-end learning in such models that approximate discret e maximum-weight matching using the continuous Sinkhorn operator. Sinkhorn iter ation is attractive because it functions as a simple, easy-to-implement analog o f the softmax operator. With this, we can define the Gumbel-Sinkhorn method, an extension of the Gumbel-Softmax method (Jang et al. 2016, Maddison2016 et al. 20 16) to distributions over latent matchings. We demonstrate the effectiveness of our method by outperforming competitive baselines on a range of qualitatively di fferent tasks: sorting numbers, solving jigsaw puzzles, and identifying neural s ignals in worms.
****
GATED FAST WEIGHTS FOR ASSOCIATIVE RETRIEVAL
Imanol Schlag,Jürgen Schmidhuber
We improve previous end-to-end differentiable neural networks (NNs) with fast
weight memories. A gate mechanism updates fast weights at every time step of
a sequence through two separate outer-product-based matrices generated by slow
parts of the net. The system is trained on a complex sequence to sequence variat ion
of the Associative Retrieval Problem with roughly 70 times more temporal
memory (i.e. time-varying variables) than similar-sized standard recurrent NNs
(RNNs). In terms of accuracy and number of parameters, our architecture outperfo rms
a variety of RNNs, including Long Short-Term Memory, Hypernetworks,
and related fast weight architectures.
****
A Framework for the Quantitative Evaluation of Disentangled Representations
Cian Eastwood,Christopher K. I. Williams
Recent AI research has emphasised the importance of learning disentangled repres entations of the explanatory factors behind data. Despite the growing interest in models which can learn such representations, visual inspection remains the st andard evaluation metric. While various desiderata have been implied in recent d efinitions, it is currently unclear what exactly makes one disentangled represen tation better than another. In this work we propose a framework for the quantita tive evaluation of disentangled representations when the ground-truth latent str ucture is available. Three criteria are explicitly defined and quantified to elu cidate the quality of learnt representations and thus compare models on an equal basis. To illustrate the appropriateness of the framework, we employ it to comp are quantitatively the representations learned by recent state-of-the-art models .
****
Long Short-Term Memory as a Dynamically Computed Element-wise Weighted Sum
Omer Levy,Kenton Lee,Nicholas FitzGerald,Luke Zettlemoyer
Long short-term memory networks (LSTMs) were introduced to combat vanishing grad ients in simple recurrent neural networks (S-RNNs) by augmenting them with addit ive recurrent connections controlled by gates. We present an alternate view to e xplain the success of LSTMs: the gates themselves are powerful recurrent models that provide more representational power than previously appreciated. We do this by showing that the LSTM's gates can be decoupled from the embedded S-RNN, prod ucing a restricted class of RNNs where the main recurrence computes an element-w ise weighted sum of context-independent functions of the inputs. Experiments on a range of challenging NLP problems demonstrate that the simplified gate-based m odels work substantially better than S-RNNs, and often just as well as the origi nal LSTMs, strongly suggesting that the gates are doing much more in practice th an just alleviating vanishing gradients.

****
Toward learning better metrics for sequence generation training with policy gradient
Joji Toyama,Yusuke Iwasawa,Kotaro Nakayama,Yutaka Matsuo
Designing a metric manually for unsupervised sequence generation tasks, such as text generation, is essentially difficult. In a such situation, learning a metric of a sequence from data is one possible solution. The previous study, SeqGAN, proposed the framework for unsupervised sequence generation, in which a metric is learned from data, and a generator is optimized with regard to the learned metric with policy gradient, inspired by generative adversarial nets (GANs) and reinforcement learning. In this paper, we make two proposals to learn better metric than SeqGAN's: partial reward function and expert-based reward function training. The partial reward function is a reward function for a partial sequence of a certain length. SeqGAN employs a reward function for completed sequence only. By combining long-scale and short-scale partial reward functions, we expect a learned metric to be able to evaluate a partial correctness as well as a coherence of a sequence, as a whole. In expert-based reward function training, a reward function is trained to discriminate between an expert (or true) sequence and a fake sequence that is produced by editing an expert sequence. Expert-based reward function training is not a kind of GAN frameworks. This makes the optimization of the generator easier. We examine the effect of the partial reward function and expert-based reward function training on synthetic data and real text data, and show improvements over SeqGAN and the model trained with MLE. Specifically, whereas SeqGAN gains 0.42 improvement of NLL over MLE on synthetic data, our best model gains 3.02 improvement, and whereas SeqGAN gains 0.029 improvement of BLEU over MLE, our best model gains 0.250 improvement.
****
A Neural Representation of Sketch Drawings
David Ha,Douglas Eck
We present sketch-rnn, a recurrent neural network able to construct stroke-based drawings of common objects. The model is trained on a dataset of human-drawn images representing many different classes. We outline a framework for conditional and unconditional sketch generation, and describe new robust training methods for generating coherent sketch drawings in a vector format.
****
Polar Transformer Networks
Carlos Esteves,Christine Allen-Blanchette,Xiaowei Zhou,Kostas Daniilidis
Convolutional neural networks (CNNs) are inherently equivariant to translation. Efforts to embed other forms of equivariance have concentrated solely on rotation. We expand the notion of equivariance in CNNs through the Polar Transformer Network (PTN). PTN combines ideas from the Spatial Transformer Network (STN) and canonical coordinate representations. The result is a network invariant to translation and equivariant to both rotation and scale. PTN is trained end-to-end and composed of three distinct stages: a polar origin predictor, the newly introduced polar transformer module and a classifier. PTN achieves state-of-the-art on rotated MNIST and the newly introduced SIM2MNIST dataset, an MNIST variation obtained by adding clutter and perturbing digits with translation, rotation and scaling. The ideas of PTN are extensible to 3D which we demonstrate through the Cylindrical Transformer Network.
****
Neural Speed Reading via Skim-RNN
Minjoon Seo,Sewon Min,Ali Farhadi,Hannaneh Hajishirzi
Inspired by the principles of speed reading, we introduce Skim-RNN, a recurrent neural network (RNN) that dynamically decides to update only a small fraction of the hidden state for relatively unimportant input tokens. Skim-RNN gives a significant computational advantage over an RNN that always updates the entire hidden state. Skim-RNN uses the same input and output interfaces as a standard RNN and can be easily used instead of RNNs in existing models. In our experiments, we show that Skim-RNN can achieve significantly reduced computational cost without losing accuracy compared to standard RNNs across five different natural languag

e tasks. In addition, we demonstrate that the trade-off between accuracy and spe ed of Skim-RNN can be dynamically controlled during inference time in a stable m anner. Our analysis also shows that Skim-RNN running on a single CPU offers lowe r latency compared to standard RNNs on GPUs.
****

Time-Dependent Representation for Neural Event Sequence Prediction
Yang Li,Nan Du,Samy Bengio
Existing sequence prediction methods are mostly concerned with time-independent sequences, in which the actual time span between events is irrelevant and the di stance between events is simply the difference between their order positions in the sequence. While this time-independent view of sequences is applicable for da ta such as natural languages, e.g., dealing with words in a sentence, it is inap propriate and inefficient for many real world events that are observed and colle cted at unequally spaced points of time as they naturally arise, e.g., when a pe rson goes to a grocery store or makes a phone call. The time span between events can carry important information about the sequence dependence of human behavior s. In this work, we propose a set of methods for using time in sequence predicti on. Because neural sequence models such as RNN are more amenable for handling to ken-like input, we propose two methods for time-dependent event representation, based on the intuition on how time is tokenized in everyday life and previous wo rk on embedding contextualization. We also introduce two methods for using next event duration as regularization for training a sequence prediction model. We di scuss these methods based on recurrent neural nets. We evaluate these methods as well as baseline models on five datasets that resemble a variety of sequence pr ediction tasks. The experiments revealed that the proposed methods offer accurac y gain over baseline models in a range of settings.
****

Towards Unsupervised Classification with Deep Generative Models
Dimitris Kalatzis,Konstantia Kotta,Ilias Kalamaras,Anastasios Vafeiadis,Andrew R awstron,Dimitris Tzovaras,Kostas Stamatopoulos
Deep generative models have advanced the state-of-the-art in semi-supervised cla ssification, however their capacity for deriving useful discriminative features in a completely unsupervised fashion for classification in difficult real-world data sets, where adequate manifold separation is required has not been adequatel y explored. Most methods rely on defining a pipeline of deriving features via ge nerative modeling and then applying clustering algorithms, separating the modeli ng and discriminative processes. We propose a deep hierarchical generative model which uses a mixture of discrete and continuous distributions to learn to effec tively separate the different data manifolds and is trainable end-to-end. We sho w that by specifying the form of the discrete variable distribution we are impos ing a specific structure on the model's latent representations. We test our mode l's discriminative performance on the task of CLL diagnosis against baselines fr om the field of computational FC, as well as the Variational Autoencoder literat ure.
****

QANet: Combining Local Convolution with Global Self-Attention for Reading Compre hension
Adams Wei Yu,David Dohan,Minh-Thang Luong,Rui Zhao,Kai Chen,Mohammad Norouzi,Quo c V. Le
Current end-to-end machine reading and question answering (Q\&A) models are pri marily based on recurrent neural networks (RNNs) with attention. Despite their s uccess, these models are often slow for both training and inference due to the s equential nature of RNNs. We propose a new Q\&A architecture called QANet, which does not require recurrent networks: Its encoder consists exclusively of convo lution and self-attention, where convolution models local interactions and self- attention models global interactions. On the SQuAD dataset, our model is 3x to 1 3x faster in training and 4x to 9x faster in inference, while achieving equivale nt accuracy to recurrent models. The speed-up gain allows us to train the model with much more data. We hence combine our model with data generated by backtrans lation from a neural machine translation model.

On the SQuAD dataset, our single model, trained with augmented data, achieves 84
.6 F1 score on the test set, which is significantly better than the best publish
ed F1 score of 81.8.
****
Deep Bayesian Bandits Showdown: An Empirical Comparison of Bayesian Deep Networks for Thompson Sampling
Carlos Riquelme,George Tucker,Jasper Snoek
Recent advances in deep reinforcement learning have made significant strides in performance on applications such as Go and Atari games. However, developing practical methods to balance exploration and exploitation in complex domains remains largely unsolved. Thompson Sampling and its extension to reinforcement learning provide an elegant approach to exploration that only requires access to posterior samples of the model. At the same time, advances in approximate Bayesian methods have made posterior approximation for flexible neural network models practical. Thus, it is attractive to consider approximate Bayesian neural networks in a Thompson Sampling framework. To understand the impact of using an approximate posterior on Thompson Sampling, we benchmark well-established and recently developed methods for approximate posterior sampling combined with Thompson Sampling over a series of contextual bandit problems. We found that many approaches that have been successful in the supervised learning setting underperformed in the sequential decision-making scenario. In particular, we highlight the challenge of adapting slowly converging uncertainty estimates to the online setting.
****
Relational Multi-Instance Learning for Concept Annotation from Medical Time Series
Sanjay Purushotham,Zhengping Che,Bo Jiang,Tanachat Nilanon,Yan Liu
Recent advances in computing technology and sensor design have made it easier to collect longitudinal or time series data from patients, resulting in a gigantic amount of available medical data. Most of the medical time series lack annotations or even when the annotations are available they could be subjective and prone to human errors. Earlier works have developed natural language processing techniques to extract concept annotations and/or clinical narratives from doctor notes. However, these approaches are slow and do not use the accompanying medical time series data. To address this issue, we introduce the problem of concept annotation for the medical time series data, i.e., the task of predicting and localizing medical concepts by using the time series data as input. We propose Relational Multi-Instance Learning (RMIL) - a deep Multi Instance Learning framework based on recurrent neural networks, which uses pooling functions and attention mechanisms for the concept annotation tasks. Empirical results on medical datasets show that our proposed models outperform various multi-instance learning models.
****
The power of deeper networks for expressing natural functions
David Rolnick,Max Tegmark
It is well-known that neural networks are universal approximators, but that deeper networks tend in practice to be more powerful than shallower ones. We shed light on this by proving that the total number of neurons m required to approximate natural classes of multivariate polynomials of n variables grows only linearly with n for deep neural networks, but grows exponentially when merely a single hidden layer is allowed. We also provide evidence that when the number of hidden layers is increased from 1 to k, the neuron requirement grows exponentially not with n but with n^{1/k}, suggesting that the minimum number of layers required for practical expressibility grows only logarithmically with n.
****
Learning to Infer
Joseph Marino,Yisong Yue,Stephan Mandt
Inference models, which replace an optimization-based inference procedure with a learned model, have been fundamental in advancing Bayesian deep learning, the most notable example being variational auto-encoders (VAEs). In this paper, we propose iterative inference models, which learn how to optimize a variational lower bound through repeatedly encoding gradients. Our approach generalizes VAEs und

er certain conditions, and by viewing VAEs in the context of iterative inference, we provide further insight into several recent empirical findings. We demonstrate the inference optimization capabilities of iterative inference models, explore unique aspects of these models, and show that they outperform standard inference models on typical benchmark data sets.
****

Generalized Graph Embedding Models
Qiao Liu,Xiaohui Yang,Rui Wan,Shouzhong Tu,Zufeng Wu
Many types of relations in physical, biological, social and information systems can be modeled as homogeneous or heterogeneous concept graphs. Hence, learning from and with graph embeddings has drawn a great deal of research interest recently, but only ad hoc solutions have been obtained this far. In this paper, we conjecture that the one-shot supervised learning mechanism is a bottleneck in improving the performance of the graph embedding learning algorithms, and propose to extend this by introducing a multi-shot unsupervised learning framework. Empirical results on several real-world data set show that the proposed model consistently and significantly outperforms existing state-of-the-art approaches on knowledge base completion and graph based multi-label classification tasks.
****

DORA The Explorer: Directed Outreaching Reinforcement Action-Selection
Lior Fox,Leshem Choshen,Yonatan Loewenstein
Exploration is a fundamental aspect of Reinforcement Learning, typically implemented using stochastic action-selection. Exploration, however, can be more efficient if directed toward gaining new world knowledge. Visit-counters have been proven useful both in practice and in theory for directed exploration. However, a major limitation of counters is their locality. While there are a few model-based solutions to this shortcoming, a model-free approach is still missing.
We propose E-values, a generalization of counters that can be used to evaluate the propagating exploratory value over state-action trajectories. We compare our approach to commonly used RL techniques, and show that using E-values improves learning and performance over traditional counters. We also show how our method can be implemented with function approximation to efficiently learn continuous MDPs. We demonstrate this by showing that our approach surpasses state of the art performance in the Freeway Atari 2600 game.
****

Learning Gaussian Policies from Smoothed Action Value Functions
Ofir Nachum,Mohammad Norouzi,George Tucker,Dale Schuurmans
State-action value functions (i.e., Q-values) are ubiquitous in reinforcement learning (RL), giving rise to popular algorithms such as SARSA and Q-learning. We propose a new notion of action value defined by a Gaussian smoothed version of the expected Q-value used in SARSA. We show that such smoothed Q-values still satisfy a Bellman equation, making them naturally learnable from experience sampled from an environment. Moreover, the gradients of expected reward with respect to the mean and covariance of a parameterized Gaussian policy can be recovered from the gradient and Hessian of the smoothed Q-value function. Based on these relationships we develop new algorithms for training a Gaussian policy directly from a learned Q-value approximator. The approach is also amenable to proximal optimization techniques by augmenting the objective with a penalty on KL-divergence from a previous policy. We find that the ability to learn both a mean and covariance during training allows this approach to achieve strong results on standard continuous control benchmarks.
****

Word2net: Deep Representations of Language
Maja Rudolph,Francisco Ruiz,David Blei
Word embeddings extract semantic features of words from large datasets of text. Most embedding methods rely on a log-bilinear model to predict the occurrence of a word in a context of other words. Here we propose word2net, a method that replaces their linear parametrization with neural networks. For each term in the vocabulary, word2net posits a neural network that takes the context as input and outputs a probability of occurrence. Further, word2net can use the hierarchical

organization of its word networks to incorporate additional meta-data, such as syntactic features, into the embedding model. For example, we show how to share parameters across word networks to develop an embedding model that includes part-of-speech information. We study word2net with two datasets, a collection of Wikipedia articles and a corpus of U.S. Senate speeches. Quantitatively, we found that word2net outperforms popular embedding methods on predicting held-out words and that sharing parameters based on part of speech further boosts performance. Qualitatively, word2net learns interpretable semantic representations
and, compared to vector-based methods, better incorporates syntactic information.
****

Deep Temporal Clustering: Fully unsupervised learning of time-domain features
Naveen Sai Madiraju,Seid M. Sadat,Dimitry Fisher,Homa Karimabadi
Unsupervised learning of timeseries data is a challenging problem in machine learning. Here,
we propose a novel algorithm, Deep Temporal Clustering (DTC), a fully unsupervised method, to naturally integrate dimensionality reduction and temporal clustering into a single end to end learning framework. The algorithm starts with an initial cluster estimates using an autoencoder for dimensionality reduction and a novel temporal clustering layer for cluster assignment. Then it jointly optimizes the clustering objective and the dimensionality reduction objective. Based on requirement and application, the temporal clustering layer can be customized with any temporal similarity metric. Several similarity metrics are considered and compared. To gain insight into features that the network has learned for its clustering, we apply a visualization method that generates a heat map of regions of interest in the timeseries. The viability of the algorithm is demonstrated using timeseries data from diverse domains, ranging from earthquakes to sensor data from spacecraft. In each case, we show that our algorithm outperforms traditional methods. This performance is attributed to fully integrated temporal dimensionality reduction and clustering criterion.
****

Subspace Network: Deep Multi-Task Censored Regression for Modeling Neurodegenerative Diseases
Mengying Sun,Inci M. Baytas,Zhangyang Wang,Jiayu Zhou
Over the past decade a wide spectrum of machine learning models have been developed to model the neurodegenerative diseases, associating biomarkers, especially non-intrusive neuroimaging markers, with key clinical scores measuring the cognitive status of patients. Multi-task learning (MTL) has been extensively explored in these studies to address challenges associated to high dimensionality and small cohort size. However, most existing MTL approaches are based on linear models and suffer from two major limitations: 1) they cannot explicitly consider upper/lower bounds in these clinical scores; 2) they lack the capability to capture complicated non-linear effects among the variables. In this paper, we propose the Subspace Network, an efficient deep modeling approach for non-linear multi-task censored regression. Each layer of the subspace network performs a multi-task censored regression to improve upon the predictions from the last layer via sketching a low-dimensional subspace to perform knowledge transfer among learning tasks. We show that under mild assumptions, for each layer the parametric subspace can be recovered using only one pass of training data. In addition, empirical results demonstrate that the proposed subspace network quickly picks up correct parameter subspaces, and outperforms state-of-the-arts in predicting neurodegenerative clinical scores using information in brain imaging.
****

Egocentric Spatial Memory Network
Mengmi Zhang,Keng Teck Ma,Joo Hwee Lim,Shih-Cheng Yen,Qi Zhao,Jiashi Feng
Inspired by neurophysiological discoveries of navigation cells in the mammalian brain, we introduce the first deep neural network architecture for modeling Egocentric
Spatial Memory (ESM). It learns to estimate the pose of the agent and

progressively construct top-down 2D global maps from egocentric views in a spati
ally
extended environment. During the exploration, our proposed ESM network
model updates belief of the global map based on local observations using a recur
rent
neural network. It also augments the local mapping with a novel external
memory to encode and store latent representations of the visited places based on
their corresponding locations in the egocentric coordinate. This enables the age
nts
to perform loop closure and mapping correction. This work contributes in the
following aspects: first, our proposed ESM network provides an accurate mapping
ability which is vitally important for embodied agents to navigate to goal locat
ions.
In the experiments, we demonstrate the functionalities of the ESM network in
random walks in complicated 3D mazes by comparing with several competitive
baselines and state-of-the-art Simultaneous Localization and Mapping (SLAM)
algorithms. Secondly, we faithfully hypothesize the functionality and the workin
g
mechanism of navigation cells in the brain. Comprehensive analysis of our model
suggests the essential role of individual modules in our proposed architecture a
nd
demonstrates efficiency of communications among these modules. We hope this
work would advance research in the collaboration and communications over both
fields of computer science and computational neuroscience.
****
Learning a Generative Model for Validity in Complex Discrete Structures
Dave Janz,Jos van der Westhuizen,Brooks Paige,Matt Kusner,José Miguel Hernández-
Lobato
Deep generative models have been successfully used to learn representations for
high-dimensional discrete spaces by representing discrete objects as sequences a
nd employing powerful sequence-based deep models. Unfortunately, these sequence-
based models often produce invalid sequences: sequences which do not represent a
ny underlying discrete structure; invalid sequences hinder the utility of such m
odels. As a step towards solving this problem, we propose to learn a deep recurr
ent validator model, which can estimate whether a partial sequence can function
as the beginning of a full, valid sequence. This validator provides insight as t
o how individual sequence elements influence the validity of the overall sequenc
e, and can be used to constrain sequence based models to generate valid sequence
s — and thus faithfully model discrete objects. Our approach is inspired by rein
forcement learning, where an oracle which can evaluate validity of complete sequ
ences provides a sparse reward signal. We demonstrate its effectiveness as a gen
erative model of Python 3 source code for mathematical expressions, and in impro
ving the ability of a variational autoencoder trained on SMILES strings to decod
e valid molecular structures.
****
Progressive Reinforcement Learning with Distillation for Multi-Skilled Motion Co
ntrol
Glen Berseth,Cheng Xie,Paul Cernek,Michiel Van de Panne
Deep reinforcement learning has demonstrated increasing capabilities for continu
ous control problems,
including agents that can move with skill and agility through their environment.

An open problem in this setting is that of developing good strategies for integr
ating or merging policies
for multiple skills, where each individual skill is a specialist in a specific s
kill and its associated state distribution.
We extend policy distillation methods to the continuous action setting and lever
age this technique to combine \expert policies,
as evaluated in the domain of simulated bipedal locomotion across different clas
ses of terrain.

We also introduce an input injection method for augmenting an existing policy ne
twork to exploit new input features.
Lastly, our method uses transfer learning to assist in the efficient acquisition
 of new skills.
The combination of these methods allows a policy to be incrementally augmented w
ith new skills.
We compare our progressive learning and integration via distillation (PLAID) met
hod
against three alternative baselines.
****
Exploring the Hidden Dimension in Accelerating Convolutional Neural Networks
Zhihao Jia,Sina Lin,Charles R. Qi,Alex Aiken
DeePa is a deep learning framework that explores parallelism in all parallelizab
le dimensions to accelerate the training process of convolutional neural network
s. DeePa optimizes parallelism at the granularity of each individual layer in th
e network. We present an elimination-based algorithm that finds an optimal paral
lelism configuration for every layer. Our evaluation shows that DeePa achieves u
p to 6.5× speedup compared to state-of-the-art deep learning frameworks and redu
ces data transfers by up to 23×.
****
Expressive power of recurrent neural networks
Valentin Khrulkov,Alexander Novikov,Ivan Oseledets
Deep neural networks are surprisingly efficient at solving practical tasks,
but the theory behind this phenomenon is only starting to catch up with
the practice. Numerous works show that depth is the key to this efficiency.
A certain class of deep convolutional networks – namely those that correspond
to the Hierarchical Tucker (HT) tensor decomposition – has been
proven to have exponentially higher expressive power than shallow networks.
I.e. a shallow network of exponential width is required to realize
the same score function as computed by the deep architecture. In this paper,
we prove the expressive power theorem (an exponential lower bound on
the width of the equivalent shallow network) for a class of recurrent neural
networks – ones that correspond to the Tensor Train (TT) decomposition.
This means that even processing an image patch by patch with an RNN
can be exponentially more efficient than a (shallow) convolutional network
with one hidden layer. Using theoretical results on the relation between
the tensor decompositions we compare expressive powers of the HT- and
TT-Networks. We also implement the recurrent TT-Networks and provide
numerical evidence of their expressivity.
****
SQLNet: Generating Structured Queries From Natural Language Without Reinforcemen
t Learning
Xiaojun Xu,Chang Liu,Dawn Song
Synthesizing SQL queries from natural language is a long-standing open problem a
nd has been attracting considerable interest recently. Toward solving the proble
m, the de facto approach is to employ a sequence-to-sequence-style model. Such a
n approach will necessarily require the SQL queries to be serialized. Since the
same SQL query may have multiple equivalent serializations, training a sequence-
to-sequence-style model is sensitive to the choice from one of them. This phenom
enon is documented as the "order-matters" problem. Existing state-of-the-art app
roaches rely on reinforcement learning to reward the decoder when it generates a
ny of the equivalent serializations. However, we observe that the improvement fr
om reinforcement learning is limited.

In this paper, we propose a novel approach, i.e., SQLNet, to fundamentally solve
 this problem by avoiding the sequence-to-sequence structure when the order does
 not matter. In particular, we employ a sketch-based approach where the sketch c
ontains a dependency graph, so that one prediction can be done by taking into co
nsideration only the previous predictions that it depends on. In addition, we pr
opose a sequence-to-set model as well as the column attention mechanism to synth

esize the query based on the sketch. By combining all these novel techniques, we show that SQLNet can outperform the prior art by 9% to 13% on the WikiSQL task.
****
Training wide residual networks for deployment using a single bit for each weight
Mark D. McDonnell
For fast and energy-efficient deployment of trained deep neural networks on resource-constrained embedded hardware, each learned weight parameter should ideally be represented and stored using a single bit. Error-rates usually increase when this requirement is imposed. Here, we report large improvements in error rates on multiple datasets, for deep convolutional neural networks deployed with 1-bit-per-weight. Using wide residual networks as our main baseline, our approach simplifies existing methods that binarize weights by applying the sign function in training; we apply scaling factors for each layer with constant unlearned values equal to the layer-specific standard deviations used for initialization. For CIFAR-10, CIFAR-100 and ImageNet, and models with 1-bit-per-weight requiring less than 10 MB of parameter memory, we achieve error rates of 3.9%, 18.5% and 26.0% / 8.5% (Top-1 / Top-5) respectively. We also considered MNIST, SVHN and ImageNet32, achieving 1-bit-per-weight test results of 0.27%, 1.9%, and 41.3% / 19.1% respectively. For CIFAR, our error rates halve previously reported values, and are within about 1% of our error-rates for the same network with full-precision weights. For networks that overfit, we also show significant improvements in error rate by not learning batch normalization scale and offset parameters. This applies to both full precision and 1-bit-per-weight networks. Using a warm-restart learning-rate schedule, we found that training for 1-bit-per-weight is just as fast as full-precision networks, with better accuracy than standard schedules, and achieved about 98%-99% of peak performance in just 62 training epochs for CIFAR-10/100. For full training code and trained models in MATLAB, Keras and PyTorch see https://github.com/McDonnell-Lab/1-bit-per-weight/ .
****
Dependent Bidirectional RNN with Extended-long Short-term Memory
Yuanhang Su,Yuzhong Huang,C.-C. Jay Kuo
In this work, we first conduct mathematical analysis on the memory, which is defined as a function that maps an element in a sequence to the current output, of three RNN cells; namely, the simple recurrent neural network (SRN), the long short-term memory (LSTM) and the gated recurrent unit (GRU). Based on the analysis, we propose a new design, called the extended-long short-term memory (ELSTM), to extend the memory length of a cell. Next, we present a multi-task RNN model that is robust to previous erroneous predictions, called the dependent bidirectional recurrent neural network (DBRNN), for the sequence-in-sequenceout (SISO) problem. Finally, the performance of the DBRNN model with the ELSTM cell is demonstrated by experimental results.
****
Natural Language Inference with External Knowledge
Qian Chen,Xiaodan Zhu,Zhen-Hua Ling,Diana Inkpen
Modeling informal inference in natural language is very challenging. With the recent availability of large annotated data, it has become feasible to train complex models such as neural networks to perform natural language inference (NLI), which have achieved state-of-the-art performance. Although there exist relatively large annotated data, can machines learn all knowledge needed to perform NLI from the data? If not, how can NLI models benefit from external knowledge and how to build NLI models to leverage it? In this paper, we aim to answer these questions by enriching the state-of-the-art neural natural language inference models with external knowledge. We demonstrate that the proposed models with external knowledge further improve the state of the art on the Stanford Natural Language Inference (SNLI) dataset.
****
Gaussian Prototypical Networks for Few-Shot Learning on Omniglot
Stanislav Fort
We propose a novel architecture for k-shot classification on the Omniglot datase

t. Building on prototypical networks, we extend their architecture to what we ca
ll Gaussian prototypical networks. Prototypical networks learn a map between ima
ges and embedding vectors, and use their clustering for classification. In our m
odel, a part of the encoder output is interpreted as a confidence region estimat
e about the embedding point, and expressed as a Gaussian covariance matrix. Our
network then constructs a direction and class dependent distance metric on the e
mbedding space, using uncertainties of individual data points as weights. We sho
w that Gaussian prototypical networks are a preferred architecture over vanilla
prototypical networks with an equivalent number of parameters. We report results
 consistent with state-of-the-art performance in 1-shot and 5-shot classificatio
n both in 5-way and 20-way regime on the Omniglot dataset. We explore artificial
ly down-sampling a fraction of images in the training set, which improves our pe
rformance. Our experiments therefore lead us to hypothesize that Gaussian protot
ypical networks might perform better in less homogeneous, noisier datasets, whic
h are commonplace in real world applications.
****
Learning Discrete Weights Using the Local Reparameterization Trick
Oran Shayer,Dan Levi,Ethan Fetaya
Recent breakthroughs in computer vision make use of large deep neural networks,
utilizing the substantial speedup offered by GPUs. For applications running on l
imited hardware, however, high precision real-time processing can still be a cha
llenge. One approach to solving this problem is training networks with binary o
r ternary weights, thus removing the need to calculate multiplications and signi
ficantly reducing memory size. In this work, we introduce LR-nets (Local reparam
eterization networks), a new method for training neural networks with discrete w
eights using stochastic parameters. We show how a simple modification to the loc
al reparameterization trick, previously used to train Gaussian distributed weigh
ts, enables the training of discrete weights. Using the proposed training we tes
t both binary and ternary models on MNIST, CIFAR-10 and ImageNet benchmarks and
reach state-of-the-art results on most experiments.
****
Can Neural Networks Understand Logical Entailment?
Richard Evans,David Saxton,David Amos,Pushmeet Kohli,Edward Grefenstette
We introduce a new dataset of logical entailments for the purpose of measuring m
odels' ability to capture and exploit the structure of logical expressions again
st an entailment prediction task. We use this task to compare a series of archit
ectures which are ubiquitous in the sequence-processing literature, in addition
to a new model class---PossibleWorldNets---which computes entailment as a ``conv
olution over possible worlds''. Results show that convolutional networks present
 the wrong inductive bias for this class of problems relative to LSTM RNNs, tree
-structured neural networks outperform LSTM RNNs due to their enhanced ability t
o exploit the syntax of logic, and PossibleWorldNets outperform all benchmarks.
****
Scalable Private Learning with PATE
Nicolas Papernot,Shuang Song,Ilya Mironov,Ananth Raghunathan,Kunal Talwar,Ulfar
Erlingsson
The rapid adoption of machine learning has increased concerns about the privacy
implications of machine learning models trained on sensitive data, such as medic
al records or other personal information. To address those concerns, one promisi
ng approach is Private Aggregation of Teacher Ensembles, or PATE, which transfer
s to a "student" model the knowledge of an ensemble of "teacher" models, with in
tuitive privacy provided by training teachers on disjoint data and strong privac
y guaranteed by noisy aggregation of teachers’ answers. However, PATE has so far
 been evaluated only on simple classification tasks like MNIST, leaving unclear
its utility when applied to larger-scale learning tasks and real-world datasets.

In this work, we show how PATE can scale to learning tasks with large numbers of
 output classes and uncurated, imbalanced training data with errors. For this, w
e introduce new noisy aggregation mechanisms for teacher ensembles that are more
 selective and add less noise, and prove their tighter differential-privacy guar

antees. Our new mechanisms build on two insights: the chance of teacher consensus is increased by using more concentrated noise and, lacking consensus, no answer need be given to a student. The consensus answers used are more likely to be correct, offer better intuitive privacy, and incur lower-differential privacy cost. Our evaluation shows our mechanisms improve on the original PATE on all measures, and scale to larger tasks with both high utility and very strong privacy ($\varepsilon < 1.0$).
****

Autoregressive Generative Adversarial Networks
Yasin Yazici,Kim-Hui Yap,Stefan Winkler
Generative Adversarial Networks (GANs) learn a generative model by playing an adversarial game between a generator and an auxiliary discriminator, which classifies data samples vs. generated ones. However, it does not explicitly model feature co-occurrences in samples. In this paper, we propose a novel Autoregressive Generative Adversarial Network (ARGAN), that models the latent distribution of data using an autoregressive model, rather than relying on binary classification of samples into data/generated categories. In this way, feature co-occurrences in samples can be more efficiently captured. Our model was evaluated on two widely used datasets: CIFAR-10 and STL-10. Its performance is competitive with respect to other GAN models both quantitatively and qualitatively.
****

Emergent Complexity via Multi-Agent Competition
Trapit Bansal,Jakub Pachocki,Szymon Sidor,Ilya Sutskever,Igor Mordatch
Reinforcement learning algorithms can train agents that solve problems in complex, interesting environments. Normally, the complexity of the trained agent is closely related to the complexity of the environment. This suggests that a highly capable agent requires a complex environment for training. In this paper, we point out that a competitive multi-agent environment trained with self-play can produce behaviors that are far more complex than the environment itself. We also point out that such environments come with a natural curriculum, because for any skill level, an environment full of agents of this level will have the right level of difficulty.
This work introduces several competitive multi-agent environments where agents compete in a 3D world with simulated physics. The trained agents learn a wide variety of complex and interesting skills, even though the environment themselves are relatively simple. The skills include behaviors such as running, blocking, ducking, tackling, fooling opponents, kicking, and defending using both arms and legs. A highlight of the learned behaviors can be found here: https://goo.gl/eR7fbX
****

Non-Autoregressive Neural Machine Translation
Jiatao Gu,James Bradbury,Caiming Xiong,Victor O.K. Li,Richard Socher
Existing approaches to neural machine translation condition each output word on previously generated outputs. We introduce a model that avoids this autoregressive property and produces its outputs in parallel, allowing an order of magnitude lower latency during inference. Through knowledge distillation, the use of input token fertilities as a latent variable, and policy gradient fine-tuning, we achieve this at a cost of as little as 2.0 BLEU points relative to the autoregressive Transformer network used as a teacher. We demonstrate substantial cumulative improvements associated with each of the three aspects of our training strategy, and validate our approach on IWSLT 2016 English–German and two WMT language pairs. By sampling fertilities in parallel at inference time, our non-autoregressive model achieves near-state-of-the-art performance of 29.8 BLEU on WMT 2016 English–Romanian.
****

Online Hyper-Parameter Optimization
Damien Vincent,Sylvain Gelly,Nicolas Le Roux,Olivier Bousquet
We propose an efficient online hyperparameter optimization method which uses a joint dynamical system to evaluate the gradient with respect to the hyperparameters. While similar methods are usually limited to hyperparameters with a smooth i

mpact on the model, we show how to apply it to the probability of dropout in neural networks. Finally, we show its effectiveness on two distinct tasks.
****
Demystifying overcomplete nonlinear auto-encoders: fast SGD convergence towards sparse representation from random initialization
Cheng Tang,Claire Monteleoni
Auto-encoders are commonly used for unsupervised representation learning and for pre-training deeper neural networks.
When its activation function is linear and the encoding dimension (width of hidden layer) is smaller than the input dimension, it is well known that auto-encoder is optimized to learn the principal components of the data distribution (Oja1982).
However, when the activation is nonlinear and when the width is larger than the input dimension (overcomplete), auto-encoder behaves differently from PCA, and in fact is known to perform well empirically for sparse coding problems.

We provide a theoretical explanation for this empirically observed phenomenon, when rectified-linear unit (ReLu) is adopted as the activation function and the hidden-layer width is set to be large.
In this case, we show that, with significant probability, initializing the weight matrix of an auto-encoder by sampling from a spherical Gaussian distribution followed by stochastic gradient descent (SGD) training converges towards the ground-truth representation for a class of sparse dictionary learning models.
In addition, we can show that, conditioning on convergence, the expected convergence rate is O(1/t), where t is the number of updates.
Our analysis quantifies how increasing hidden layer width helps the training performance when random initialization is used, and how the norm of network weights influence the speed of SGD convergence.
****
Generative networks as inverse problems with Scattering transforms
Tomás Angles,Stéphane Mallat
Generative Adversarial Nets (GANs) and Variational Auto-Encoders (VAEs) provide impressive image generations from Gaussian white noise, but the underlying mathematics are not well understood. We compute deep convolutional network generators by inverting a fixed embedding operator. Therefore, they do not require to be optimized with a discriminator or an encoder. The embedding is Lipschitz continuous to deformations so that generators transform linear interpolations between input white noise vectors into deformations between output images. This embedding is computed with a wavelet Scattering transform. Numerical experiments demonstrate that the resulting Scattering generators have similar properties as GANs or VAEs, without learning a discriminative network or an encoder.
****
SpectralNet: Spectral Clustering using Deep Neural Networks
Uri Shaham,Kelly Stanton,Henry Li,Ronen Basri,Boaz Nadler,Yuval Kluger
Spectral clustering is a leading and popular technique in unsupervised data analysis. Two of its major limitations are scalability and generalization of the spectral embedding (i.e., out-of-sample-extension). In this paper we introduce a deep learning approach to spectral clustering that overcomes the above shortcomings. Our network, which we call SpectralNet, learns a map that embeds input data points into the eigenspace of their associated graph Laplacian matrix and subsequently clusters them. We train SpectralNet using a procedure that involves constrained stochastic optimization. Stochastic optimization allows it to scale to large datasets, while the constraints, which are implemented using a special purpose output layer, allow us to keep the network output orthogonal. Moreover, the map learned by SpectralNet naturally generalizes the spectral embedding to unseen data points. To further improve the quality of the clustering, we replace the standard pairwise Gaussian affinities with affinities leaned from unlabeled data using a Siamese network. Additional improvement can be achieved by applying the network to code representations produced, e.g., by standard autoencoders. Our end-to-end learning procedure is fully unsupervised. In addition, we apply VC dim

ension theory to derive a lower bound on the size of SpectralNet. State-of-the-art clustering results are reported for both the MNIST and Reuters datasets.

****

Learning Representations for Faster Similarity Search
Ludwig Schmidt,Kunal Talwar
In high dimensions, the performance of nearest neighbor algorithms depends crucially on structure in the data.
While traditional nearest neighbor datasets consisted mostly of hand-crafted feature vectors, an increasing number of datasets comes from representations learned with neural networks.
We study the interaction between nearest neighbor algorithms and neural networks in more detail.
We find that the network architecture can significantly influence the efficacy of nearest neighbor algorithms even when the classification accuracy is unchanged.
Based on our experiments, we propose a number of training modifications that lead to significantly better datasets for nearest neighbor algorithms.
Our modifications lead to learned representations that can accelerate nearest neighbor queries by 5x.
****

Fast Node Embeddings: Learning Ego-Centric Representations
Tiago Pimentel,Adriano Veloso,Nivio Ziviani
Representation learning is one of the foundations of Deep Learning and allowed important improvements on several Machine Learning tasks, such as Neural Machine Translation, Question Answering and Speech Recognition. Recent works have proposed new methods for learning representations for nodes and edges in graphs. Several of these methods are based on the SkipGram algorithm, and they usually process a large number of multi-hop neighbors in order to produce the context from which node representations are learned. In this paper, we propose an effective and also efficient method for generating node embeddings in graphs that employs a restricted number of permutations over the immediate neighborhood of a node as context to generate its representation, thus ego-centric representations. We present a thorough evaluation showing that our method outperforms state-of-the-art methods in six different datasets related to the problems of link prediction and node classification, being one to three orders of magnitude faster than baselines when generating node embeddings for very large graphs.
****

SGD Learns Over-parameterized Networks that Provably Generalize on Linearly Separable Data
Alon Brutzkus,Amir Globerson,Eran Malach,Shai Shalev-Shwartz
Neural networks exhibit good generalization behavior in the
over-parameterized regime, where the number of network parameters
exceeds the number of observations. Nonetheless,
current generalization bounds for neural networks fail to explain this
phenomenon. In an attempt to bridge this gap, we study the problem of
learning a two-layer over-parameterized neural network, when the data is generated by a linearly separable function. In the case where the network has Leaky ReLU activations, we provide both optimization and generalization guarantees for over-parameterized networks.
Specifically, we prove convergence rates of SGD to a global
minimum and provide generalization guarantees for this global minimum
that are independent of the network size.
Therefore, our result clearly shows that the use of SGD for optimization both finds a global minimum, and avoids overfitting despite the high capacity of the model. This is the first theoretical demonstration that SGD can avoid overfitting, when learning over-specified neural network classifiers.
****

Discovery of Predictive Representations With a Network of General Value Functions

Matthew Schlegel,Andrew Patterson,Adam White,Martha White
The ability of an agent to {\em discover} its own learning objectives has long been considered a key ingredient for artificial general intelligence. Breakthroughs in autonomous decision making and reinforcement learning have primarily been in domains where the agent's goal is outlined and clear: such as playing a game to win, or driving safely. Several studies have demonstrated that learning extramural sub-tasks and auxiliary predictions can improve (1) single human-specified task learning, (2) transfer of learning, (3) and the agent's learned representation of the world. In all these examples, the agent was instructed what to learn about. We investigate a framework for discovery: curating a large collection of predictions, which are used to construct the agent's representation of the world. Specifically, our system maintains a large collection of predictions, continually pruning and replacing predictions. We highlight the importance of considering stability rather than convergence for such a system, and develop an adaptive, regularized algorithm towards that aim. We provide several experiments in computational micro-worlds demonstrating that this simple approach can be effective for discovering useful predictions autonomously.
****
Understanding and Exploiting the Low-Rank Structure of Deep Networks
Craig Bakker,Michael J. Henry,Nathan O. Hodas
Training methods for deep networks are primarily variants on stochastic gradient descent. Techniques that use (approximate) second-order information are rarely used because of the computational cost and noise associated with those approaches in deep learning contexts. However, in this paper, we show how feedforward deep networks exhibit a low-rank derivative structure. This low-rank structure makes it possible to use second-order information without needing approximations and without incurring a significantly greater computational cost than gradient descent. To demonstrate this capability, we implement Cubic Regularization (CR) on a feedforward deep network with stochastic gradient descent and two of its variants. There, we use CR to calculate learning rates on a per-iteration basis while training on the MNIST and CIFAR-10 datasets. CR proved particularly successful in escaping plateau regions of the objective function. We also found that this approach requires less problem-specific information (e.g. an optimal initial learning rate) than other first-order methods in order to perform well.
****
Depthwise Separable Convolutions for Neural Machine Translation
Lukasz Kaiser,Aidan N. Gomez,Francois Chollet
Depthwise separable convolutions reduce the number of parameters and computation used in convolutional operations while increasing representational efficiency. They have been shown to be successful in image classification models, both in obtaining better models than previously possible for a given parameter count (the Xception architecture) and considerably reducing the number of parameters required to perform at a given level (the MobileNets family of architectures). Recently, convolutional sequence-to-sequence networks have been applied to machine translation tasks with good results. In this work, we study how depthwise separable convolutions can be applied to neural machine translation. We introduce a new architecture inspired by Xception and ByteNet, called SliceNet, which enables a significant reduction of the parameter count and amount of computation needed to obtain results like ByteNet, and, with a similar parameter count, achieves better results.
In addition to showing that depthwise separable convolutions perform well for machine translation, we investigate the architectural changes that they enable: we observe that thanks to depthwise separability, we can increase the length of convolution windows, removing the need for filter dilation. We also introduce a new super-separable convolution operation that further reduces the number of parameters and computational cost of the models.
****
Feat2Vec: Dense Vector Representation for Data with Arbitrary Features
Luis Armona,José P. González-Brenes,Ralph Edezhath
Methods that calculate dense vector representations for features in unstructured

data—such as words in a document—have proven to be very successful for knowledge representation. We study how to estimate dense representations when multiple feature types exist within a dataset for supervised learning where explicit labels are available, as well as for unsupervised learning where there are no labels. Feat2Vec calculates embeddings for data with multiple feature types enforcing that all different feature types exist in a common space. In the supervised case, we show that our method has advantages over recently proposed methods; such as enabling higher prediction accuracy, and providing a way to avoid the cold-start problem. In the unsupervised case, our experiments suggest that Feat2Vec significantly outperforms existing algorithms that do not leverage the structure of the data. We believe that we are the first to propose a method for learning unsupervised embeddings that leverage the structure of multiple feature types.
****
Super-Convergence: Very Fast Training of Residual Networks Using Large Learning Rates
Leslie N. Smith,Nicholay Topin
In this paper, we show a phenomenon, which we named ``super-convergence'', where residual networks can be trained using an order of magnitude fewer iterations than is used with standard training methods. The existence of super-convergence is relevant to understanding why deep networks generalize well. One of the key elements of super-convergence is training with cyclical learning rates and a large maximum learning rate. Furthermore, we present evidence that training with large learning rates improves performance by regularizing the network. In addition, we show that super-convergence provides a greater boost in performance relative to standard training when the amount of labeled training data is limited. We also derive a simplification of the Hessian Free optimization method to compute an estimate of the optimal learning rate. The architectures to replicate this work will be made available upon publication.

****
Investigating Human Priors for Playing Video Games
Rachit Dubey,Pulkit Agrawal,Deepak Pathak,Thomas L. Griffiths,Alexei A. Efros
What makes humans so good at solving seemingly complex video games? Unlike computers, humans bring in a great deal of prior knowledge about the world, enabling efficient decision making. This paper investigates the role of human priors for solving video games. Given a sample game, we conduct a series of ablation studies to quantify the importance of various priors. We do this by modifying the video game environment to systematically mask different types of visual information that could be used by humans as priors. We find that removal of some prior knowledge causes a drastic degradation in the speed with which human players solve the game, e.g. from 2 minutes to over 20 minutes. Furthermore, our results indicate that general priors, such as the importance of objects and visual consistency, are critical for efficient game-play.
****
Faster Distributed Synchronous SGD with Weak Synchronization
Cong Xie,Oluwasanmi O. Koyejo,Indranil Gupta
Distributed training of deep learning is widely conducted with large neural networks and large datasets. Besides asynchronous stochastic gradient descent~(SGD), synchronous SGD is a reasonable alternative with better convergence guarantees. However, synchronous SGD suffers from stragglers. To make things worse, although there are some strategies dealing with slow workers, the issue of slow servers is commonly ignored. In this paper, we propose a new parameter server~(PS) framework dealing with not only slow workers, but also slow servers by weakening the synchronization criterion. The empirical results show good performance when there are stragglers.
****
Measuring the Intrinsic Dimension of Objective Landscapes
Chunyuan Li,Heerad Farkhoor,Rosanne Liu,Jason Yosinski
Many recently trained neural networks employ large numbers of parameters to achieve good performance. One may intuitively use the number of parameters required

as a rough gauge of the difficulty of a problem. But how accurate are such notions? How many parameters are really needed? In this paper we attempt to answer this question by training networks not in their native parameter space, but instead in a smaller, randomly oriented subspace. We slowly increase the dimension of this subspace, note at which dimension solutions first appear, and define this to be the intrinsic dimension of the objective landscape. The approach is simple to implement, computationally tractable, and produces several suggestive conclusions. Many problems have smaller intrinsic dimensions than one might suspect, and the intrinsic dimension for a given dataset varies little across a family of models with vastly different sizes. This latter result has the profound implication that once a parameter space is large enough to solve a problem, extra parameters serve directly to increase the dimensionality of the solution manifold. Intrinsic dimension allows some quantitative comparison of problem difficulty across supervised, reinforcement, and other types of learning where we conclude, for example, that solving the inverted pendulum problem is 100 times easier than classifying digits from MNIST, and playing Atari Pong from pixels is about as hard as classifying CIFAR-10. In addition to providing new cartography of the objective landscapes wandered by parameterized models, the method is a simple technique for constructively obtaining an upper bound on the minimum description length of a solution. A byproduct of this construction is a simple approach for compressing networks, in some cases by more than 100 times.
****

Adaptive Memory Networks

Daniel Li,Asim Kadav

Real-world Question Answering (QA) tasks consist of thousands of words that often represent many facts and entities. Existing models based on LSTMs require a large number of parameters to support external memory and do not generalize well for long sequence inputs. Memory networks attempt to address these limitations by storing information to an external memory module but must examine all inputs in the memory. Hence, for longer sequence inputs the intermediate memory components proportionally scale in size resulting in poor inference times and high computation costs.

In this paper, we present Adaptive Memory Networks (AMN) that process input question pairs to dynamically construct a network architecture optimized for lower inference times. During inference, AMN parses input text into entities within different memory slots. However, distinct from previous approaches, AMN is a dynamic network architecture that creates variable numbers of memory banks weighted by question relevance. Thus, the decoder can select a variable number of memory banks to construct an answer using fewer banks, creating a runtime trade-off between accuracy and speed.

AMN is enabled by first, a novel bank controller that makes discrete decisions with high accuracy and second, the capabilities of a dynamic framework (such as PyTorch) that allow for dynamic network sizing and efficient variable mini-batching. In our results, we demonstrate that our model learns to construct a varying number of memory banks based on task complexity and achieves faster inference times for standard bAbI tasks, and modified bAbI tasks. We achieve state of the art accuracy over these tasks with an average 48% lower entities are examined during inference.
****

Imitation Learning from Visual Data with Multiple Intentions

Aviv Tamar,Khashayar Rohanimanesh,Yinlam Chow,Chris Vigorito,Ben Goodrich,Michael Kahane,Derik Pridmore

Recent advances in learning from demonstrations (LfD) with deep neural networks have enabled learning complex robot skills that involve high dimensional perception such as raw image inputs.
LfD algorithms generally assume learning from single task demonstrations. In practice, however, it is more efficient for a teacher to demonstrate a multitude of tasks without careful task set up, labeling, and engineering. Unfortunately in

such cases, traditional imitation learning techniques fail to represent the multi-modal nature of the data, and often result in sub-optimal behavior. In this paper we present an LfD approach for learning multiple modes of behavior from visual data. Our approach is based on a stochastic deep neural network (SNN), which represents the underlying intention in the demonstration as a stochastic activation in the network. We present an efficient algorithm for training SNNs, and for learning with vision inputs, we also propose an architecture that associates the intention with a stochastic attention module.
We demonstrate our method on real robot visual object reaching tasks, and show that
it can reliably learn the multiple behavior modes in the demonstration data. Video results are available at https://vimeo.com/240212286/fd401241b9.
****
Emergent Translation in Multi-Agent Communication
Jason Lee,Kyunghyun Cho,Jason Weston,Douwe Kiela
While most machine translation systems to date are trained on large parallel corpora, humans learn language in a different way: by being grounded in an environment and interacting with other humans. In this work, we propose a communication game where two agents, native speakers of their own respective languages, jointly learn to solve a visual referential task. We find that the ability to understand and translate a foreign language emerges as a means to achieve shared goals. The emergent translation is interactive and multimodal, and crucially does not require parallel corpora, but only monolingual, independent text and corresponding images. Our proposed translation model achieves this by grounding the source and target languages into a shared visual modality, and outperforms several baselines on both word-level and sentence-level translation tasks. Furthermore, we show that agents in a multilingual community learn to translate better and faster than in a bilingual communication setting.
****
Emergence of grid-like representations by training recurrent neural networks to perform spatial localization
Christopher J. Cueva,Xue-Xin Wei
Decades of research on the neural code underlying spatial navigation have revealed a diverse set of neural response properties. The Entorhinal Cortex (EC) of the mammalian brain contains a rich set of spatial correlates, including grid cells which encode space using tessellating patterns. However, the mechanisms and functional significance of these spatial representations remain largely mysterious. As a new way to understand these neural representations, we trained recurrent neural networks (RNNs) to perform navigation tasks in 2D arenas based on velocity inputs. Surprisingly, we find that grid-like spatial response patterns emerge in trained networks, along with units that exhibit other spatial correlates, including border cells and band-like cells. All these different functional types of neurons have been observed experimentally. The order of the emergence of grid-like and border cells is also consistent with observations from developmental studies. Together, our results suggest that grid cells, border cells and others as observed in EC may be a natural solution for representing space efficiently given the predominant recurrent connections in the neural circuits.

****
Connectivity Learning in Multi-Branch Networks
Karim Ahmed,Lorenzo Torresani
While much of the work in the design of convolutional networks over the last five years has revolved around the empirical investigation of the importance of depth, filter sizes, and number of feature channels, recent studies have shown that branching, i.e., splitting the computation along parallel but distinct threads and then aggregating their outputs, represents a new promising dimension for significant improvements in performance. To combat the complexity of design choices in multi-branch architectures, prior work has adopted simple strategies, such as a fixed branching factor, the same input being fed to all parallel branches, and an additive combination of the outputs produced by all branches at aggregati

on points.

In this work we remove these predefined choices and propose an algorithm to lear n the connections between branches in the network. Instead of being chosen a pri ori by the human designer, the multi-branch connectivity is learned simultaneous ly with the weights of the network by optimizing a single loss function defined with respect to the end task. We demonstrate our approach on the problem of mult i-class image classification using four different datasets where it yields consi stently higher accuracy compared to the state-of-the-art ``ResNeXt'' multi-branc h network given the same learning capacity.
****

Characterizing Adversarial Subspaces Using Local Intrinsic Dimensionality

Xingjun Ma,Bo Li,Yisen Wang,Sarah M. Erfani,Sudanthi Wijewickrema,Grant Schoeneb eck,Dawn Song,Michael E. Houle,James Bailey

Deep Neural Networks (DNNs) have recently been shown to be vulnerable against ad versarial examples, which are carefully crafted instances that can mislead DNNs to make errors during prediction. To better understand such attacks, a character ization is needed of the properties of regions (the so-called `adversarial subsp aces') in which adversarial examples lie. We tackle this challenge by characteri zing the dimensional properties of adversarial regions, via the use of Local Int rinsic Dimensionality (LID). LID assesses the space-filling capability of the re gion surrounding a reference example, based on the distance distribution of the example to its neighbors. We first provide explanations about how adversarial pe rturbation can affect the LID characteristic of adversarial regions, and then sh ow empirically that LID characteristics can facilitate the distinction of advers arial examples generated using state-of-the-art attacks. As a proof-of-concept, we show that a potential application of LID is to distinguish adversarial exampl es, and the preliminary results show that it can outperform several state-of-the -art detection measures by large margins for five attack strategies considered i n this paper across three benchmark datasets. Our analysis of the LID characteri stic for adversarial regions not only motivates new directions of effective adve rsarial defense, but also opens up more challenges for developing new attacks to better understand the vulnerabilities of DNNs.
****

Regularization Neural Networks via Constrained Virtual Movement Field

Zhendong Zhang,Cheolkon Jung

We provide a novel thinking of regularization neural networks. We smooth the obj ective of neural networks w.r.t small adversarial perturbations of the inputs. D ifferent from previous works, we assume the adversarial perturbations are caused by the movement field. When the magnitude of movement field approaches 0, we ca ll it virtual movement field. By introducing the movement field, we cast the pro blem of finding adversarial perturbations into the problem of finding adversaria l movement field. By adding proper geometrical constraints to the movement field , such smoothness can be approximated in closed-form by solving a min-max proble m and its geometric meaning is clear. We define the approximated smoothness as t he regularization term. We derive three regularization terms as running example s which measure the smoothness w.r.t shift, rotation and scale respectively by a dding different constraints. We evaluate our methods on synthetic data, MNIST an d CIFAR-10. Experimental results show that our proposed method can significantly improve the baseline neural networks. Compared with the state of the art regula rization methods, proposed method achieves a tradeoff between accuracy and geome trical interpretability as well as computational cost.
****

Alpha-divergence bridges maximum likelihood and reinforcement learning in neural sequence generation

Sotetsu Koyamada,Yuta Kikuchi,Atsunori Kanemura,Shin-ichi Maeda,Shin Ishii

Neural sequence generation is commonly approached by using maximum- likelihood (ML) estimation or reinforcement learning (RL). However, it is known that they ha ve their own shortcomings; ML presents training/testing discrepancy, whereas RL suffers from sample inefficiency. We point out that it is difficult to resolve a

ll of the shortcomings simultaneously because of a tradeoff between ML and RL. In order to counteract these problems, we propose an objective function for sequence generation using α-divergence, which leads to an ML-RL integrated method that exploits better parts of ML and RL. We demonstrate that the proposed objective function generalizes ML and RL objective functions because it includes both as its special cases (ML corresponds to $\alpha \to 0$ and RL to $\alpha \to 1$). We provide a proposition stating that the difference between the RL objective function and the proposed one monotonically decreases with increasing α. Experimental results on machine translation tasks show that minimizing the proposed objective function achieves better sequence generation performance than ML-based methods.
****
Towards Safe Deep Learning: Unsupervised Defense Against Generic Adversarial Attacks
Bita Darvish Rouhani,Mohammad Samragh,Tara Javidi,Farinaz Koushanfar
Recent advances in adversarial Deep Learning (DL) have opened up a new and largely unexplored surface for malicious attacks jeopardizing the integrity of autonomous DL systems. We introduce a novel automated countermeasure called Parallel Checkpointing Learners (PCL) to thwart the potential adversarial attacks and significantly improve the reliability (safety) of a victim DL model. The proposed PCL methodology is unsupervised, meaning that no adversarial sample is leveraged to build/train parallel checkpointing learners. We formalize the goal of preventing adversarial attacks as an optimization problem to minimize the rarely observed regions in the latent feature space spanned by a DL network. To solve the aforementioned minimization problem, a set of complementary but disjoint checkpointing modules are trained and leveraged to validate the victim model execution in parallel. Each checkpointing learner explicitly characterizes the geometry of the input data and the corresponding high-level data abstractions within a particular DL layer. As such, the adversary is required to simultaneously deceive all the defender modules in order to succeed. We extensively evaluate the performance of the PCL methodology against the state-of-the-art attack scenarios, including Fast-Gradient-Sign (FGS), Jacobian Saliency Map Attack (JSMA), Deepfool, and Carlini&WagnerL2 algorithm. Extensive proof-of-concept evaluations for analyzing various data collections including MNIST, CIFAR10, and ImageNet corroborate the effectiveness of our proposed defense mechanism against adversarial samples.
****
Small Coresets to Represent Large Training Data for Support Vector Machines
Cenk Baykal,Murad Tukan,Dan Feldman,Daniela Rus
Support Vector Machines (SVMs) are one of the most popular algorithms for classification and regression analysis. Despite their popularity, even efficient implementations have proven to be computationally expensive to train at a large-scale, especially in streaming settings. In this paper, we propose a novel coreset construction algorithm for efficiently generating compact representations of massive data sets to speed up SVM training. A coreset is a weighted subset of the original data points such that SVMs trained on the coreset are provably competitive with those trained on the original (massive) data set. We provide both lower and upper bounds on the number of samples required to obtain accurate approximations to the SVM problem as a function of the complexity of the input data. Our analysis also establishes sufficient conditions on the existence of sufficiently compact and representative coresets for the SVM problem. We empirically evaluate the practical effectiveness of our algorithm against synthetic and real-world data sets.
****
Gated ConvNets for Letter-Based ASR
Vitaliy Liptchinsky,Gabriel Synnaeve,Ronan Collobert
In this paper we introduce a new speech recognition system, leveraging a simple letter-based ConvNet acoustic model. The acoustic model requires only audio transcription for training -- no alignment annotations, nor any forced alignment step is needed. At inference, our decoder takes only a word list and a language model, and is fed with letter scores from the acoustic model -- no phonetic word lexicon is needed. Key ingredients for the acoustic model are Gated Linear Units a

nd high dropout. We show near state-of-the-art results in word error rate on the LibriSpeech corpus with MFSC features, both on the clean and other configurations.

****

Data augmentation instead of explicit regularization

Alex Hernández-García,Peter König

Modern deep artificial neural networks have achieved impressive results through models with very large capacity---compared to the number of training examples---that control overfitting with the help of different forms of regularization. Regularization can be implicit, as is the case of stochastic gradient descent or parameter sharing in convolutional layers, or explicit. Most common explicit regularization techniques, such as dropout and weight decay, reduce the effective capacity of the model and typically require the use of deeper and wider architectures to compensate for the reduced capacity. Although these techniques have been proven successful in terms of results, they seem to waste capacity. In contrast, data augmentation techniques reduce the generalization error by increasing the number of training examples and without reducing the effective capacity. In this paper we systematically analyze the effect of data augmentation on some popular architectures and conclude that data augmentation alone---without any other explicit regularization techniques---can achieve the same performance or higher as regularized models, especially when training with fewer examples.

****

From Information Bottleneck To Activation Norm Penalty

Allen Nie,Mihir Mongia,James Zou

Many regularization methods have been proposed to prevent overfitting in neural networks. Recently, a regularization method has been proposed to optimize the variational lower bound of the Information Bottleneck Lagrangian. However, this method cannot be generalized to regular neural network architectures. We present the activation norm penalty that is derived from the information bottleneck principle and is theoretically grounded in a variation dropout framework. Unlike in previous literature, it can be applied to any general neural network. We demonstrate that this penalty can give consistent improvements to different state of the art architectures both in language modeling and image classification. We present analyses on the properties of this penalty and compare it to other methods that also reduce mutual information.

****

DLVM: A modern compiler infrastructure for deep learning systems

Richard Wei,Lane Schwartz,Vikram Adve

Deep learning software demands reliability and performance. However, many of the existing deep learning frameworks are software libraries that act as an unsafe DSL in Python and a computation graph interpreter. We present DLVM, a design and implementation of a compiler infrastructure with a linear algebra intermediate representation, algorithmic differentiation by adjoint code generation, domain-specific optimizations and a code generator targeting GPU via LLVM. Designed as a modern compiler infrastructure inspired by LLVM, DLVM is more modular and more generic than existing deep learning compiler frameworks, and supports tensor DSLs with high expressivity. With our prototypical staged DSL embedded in Swift, we argue that the DLVM system enables a form of modular, safe and performant frameworks for deep learning.

****

Memory Augmented Control Networks

Arbaaz Khan,Clark Zhang,Nikolay Atanasov,Konstantinos Karydis,Vijay Kumar,Daniel D. Lee

Planning problems in partially observable environments cannot be solved directly with convolutional networks and require some form of memory. But, even memory networks with sophisticated addressing schemes are unable to learn intelligent reasoning satisfactorily due to the complexity of simultaneously learning to access memory and plan. To mitigate these challenges we propose the Memory Augmented Control Network (MACN). The network splits planning into a hierarchical process.

At a lower level, it learns to plan in a locally observed space. At a higher level, it uses a collection of policies computed on locally observed spaces to learn an optimal plan in the global environment it is operating in. The performance of the network is evaluated on path planning tasks in environments in the presence of simple and complex obstacles and in addition, is tested for its ability to generalize to new environments not seen in the training set.
****
Multi-label Learning for Large Text Corpora using Latent Variable Model with Provable Gurantees
Sayantan Dasgupta
Here we study the problem of learning labels for large text corpora where each document can be assigned a variable number of labels. The problem is trivial when the label dimensionality is small and can be easily solved by a series of one-vs-all classifiers. However, as the label dimensionality increases, the parameter space of such one-vs-all classifiers becomes extremely large and outstrips the memory. Here we propose a latent variable model to reduce the size of the parameter space, but still efficiently learn the labels. We learn the model using spectral learning and show how to extract the parameters using only three passes through the training dataset. Further, we analyse the sample complexity of our model using PAC learning theory and then demonstrate the performance of our algorithm on several benchmark datasets in comparison with existing algorithms.

****
Capturing Human Category Representations by Sampling in Deep Feature Spaces
Joshua Peterson,Krishan Aghi,Jordan Suchow,Alexander Ku,Tom Griffiths
Understanding how people represent categories is a core problem in cognitive science, with the flexibility of human learning remaining a gold standard to which modern artificial intelligence and machine learning aspire. Decades of psychological research have yielded a variety of formal theories of categories, yet validating these theories with naturalistic stimuli remains a challenge. The problem is that human category representations cannot be directly observed and running informative experiments with naturalistic stimuli such as images requires having a workable representation of these stimuli. Deep neural networks have recently been successful in a range of computer vision tasks and provide a way to represent the features of images. In this paper, we introduce a method for estimating the structure of human categories that draws on ideas from both cognitive science and machine learning, blending human-based algorithms with state-of-the-art deep representation learners. We provide qualitative and quantitative results as a proof of concept for the feasibility of the method. Samples drawn from human distributions rival the quality of current state-of-the-art generative models and outperform alternative methods for estimating the structure of human categories.
****
The High-Dimensional Geometry of Binary Neural Networks
Alexander G. Anderson,Cory P. Berg
Recent research has shown that one can train a neural network with binary weights and activations at train time by augmenting the weights with a high-precision continuous latent variable that accumulates small changes from stochastic gradient descent. However, there is a dearth of work to explain why one can effectively capture the features in data with binary weights and activations. Our main result is that the neural networks with binary weights and activations trained using the method of Courbariaux, Hubara et al. (2016) work because of the high-dimensional geometry of binary vectors. In particular, the ideal continuous vectors that extract out features in the intermediate representations of these BNNs are well-approximated by binary vectors in the sense that dot products are approximately preserved. Compared to previous research that demonstrated good classification performance with BNNs, our work explains why these BNNs work in terms of HD geometry. Furthermore, the results and analysis used on BNNs are shown to generalize to neural networks with ternary weights and activations. Our theory serves as a foundation for understanding not only BNNs but a variety of methods that seek to compress traditional neural networks. Furthermore, a better understanding

of multilayer binary neural networks serves as a starting point for generalizing BNNs to other neural network architectures such as recurrent neural networks.
****
A Simple Neural Attentive Meta-Learner
Nikhil Mishra,Mostafa Rohaninejad,Xi Chen,Pieter Abbeel
Deep neural networks excel in regimes with large amounts of data, but tend to struggle when data is scarce or when they need to adapt quickly to changes in the task. In response, recent work in meta-learning proposes training a meta-learner on a distribution of similar tasks, in the hopes of generalization to novel but related tasks by learning a high-level strategy that captures the essence of the problem it is asked to solve. However, many recent meta-learning approaches are extensively hand-designed, either using architectures specialized to a particular application, or hard-coding algorithmic components that constrain how the meta-learner solves the task. We propose a class of simple and generic meta-learner architectures that use a novel combination of temporal convolutions and soft attention; the former to aggregate information from past experience and the latter to pinpoint specific pieces of information. In the most extensive set of meta-learning experiments to date, we evaluate the resulting Simple Neural AttentIve Learner (or SNAIL) on several heavily-benchmarked tasks. On all tasks, in both supervised and reinforcement learning, SNAIL attains state-of-the-art performance by significant margins.
****
Improving Deep Learning by Inverse Square Root Linear Units (ISRLUs)
Brad Carlile,Guy Delamarter,Paul Kinney,Akiko Marti,Brian Whitney
We introduce the “inverse square root linear unit” (ISRLU) to speed up learning in deep neural networks. ISRLU has better performance than ELU but has many of the same benefits. ISRLU and ELU have similar curves and characteristics. Both have negative values, allowing them to push mean unit activation closer to zero, and bring the normal gradient closer to the unit natural gradient, ensuring a noise- robust deactivation state, lessening the over fitting risk. The significant performance advantage of ISRLU on traditional CPUs also carry over to more efficient HW implementations on HW/SW codesign for CNNs/RNNs. In experiments with TensorFlow, ISRLU leads to faster learning and better generalization than ReLU on CNNs. This work also suggests a computationally efficient variant called the “inverse square root unit” (ISRU) which can be used for RNNs. Many RNNs use either long short-term memory (LSTM) and gated recurrent units (GRU) which are implemented with tanh and sigmoid activation functions. ISRU has less computational complexity but still has a similar curve to tanh and sigmoid.
****
COLD FUSION: TRAINING SEQ2SEQ MODELS TOGETHER WITH LANGUAGE MODELS
Anuroop Sriram,Heewoo Jun,Sanjeev Satheesh,Adam Coates
Sequence-to-sequence (Seq2Seq) models with attention have excelled at tasks which involve generating natural language sentences such as machine translation, image captioning and speech recognition. Performance has further been improved by leveraging unlabeled data, often in the form of a language model. In this work, we present the Cold Fusion method, which leverages a pre-trained language model during training, and show its effectiveness on the speech recognition task. We show that Seq2Seq models with Cold Fusion are able to better utilize language information enjoying i) faster convergence and better generalization, and ii) almost complete transfer to a new domain while using less than 10% of the labeled training data.
****
Deep Mean Field Theory: Layerwise Variance and Width Variation as Methods to Control Gradient Explosion
Greg Yang,Sam S. Schoenholz
■A recent line of work has studied the statistical properties of neural networks to great success from a {\it mean field theory} perspective, making and verifying very precise predictions of neural network behavior and test time performance.
■In this paper, we build upon these works to explore two methods for taming the

behaviors of random residual networks (with only fully connected layers and no b atchnorm).
■The first method is {\it width variation (WV)}, i.e. varying the widths of laye rs as a function of depth.
■We show that width decay reduces gradient explosion without affecting the mean forward dynamics of the random network.
■The second method is {\it variance variation (VV)}, i.e. changing the initializ ation variances of weights and biases over depth.
■We show VV, used appropriately, can reduce gradient explosion of tanh and ReLU resnets from $\exp(\Theta(\sqrt L))$ and $\exp(\Theta(L))$ respectively to const ant $\Theta(1)$.
■A complete phase-diagram is derived for how variance decay affects different dy namics, such as those of gradient and activation norms.
■In particular, we show the existence of many phase transitions where these dyna mics switch between exponential, polynomial, logarithmic, and even constant beha viors.
■Using the obtained mean field theory, we are able to track surprisingly well ho w VV at initialization time affects training and test time performance on MNIST after a set number of epochs: the level sets of test/train set accuracies coinci de with the level sets of the expectations of certain gradient norms or of metri c expressivity (as defined in \cite{yang_meanfield_2017}), a measure of expansio n in a random neural network.
■Based on insights from past works in deep mean field theory and information geo metry, we also provide a new perspective on the gradient explosion/vanishing pro blems: they lead to ill-conditioning of the Fisher information matrix, causing o ptimization troubles.
****
Semi-parametric topological memory for navigation
Nikolay Savinov,Alexey Dosovitskiy,Vladlen Koltun
We introduce a new memory architecture for navigation in previously unseen envir onments, inspired by landmark-based navigation in animals. The proposed semi-par ametric topological memory (SPTM) consists of a (non-parametric) graph with node s corresponding to locations in the environment and a (parametric) deep network capable of retrieving nodes from the graph based on observations. The graph stor es no metric information, only connectivity of locations corresponding to the no des. We use SPTM as a planning module in a navigation system. Given only 5 minut es of footage of a previously unseen maze, an SPTM-based navigation agent can bu ild a topological map of the environment and use it to confidently navigate towa rds goals. The average success rate of the SPTM agent in goal-directed navigatio n across test environments is higher than the best-performing baseline by a fact or of three.
****
Hierarchical Density Order Embeddings
Ben Athiwaratkun,Andrew Gordon Wilson
By representing words with probability densities rather than point vectors, prob a- bilistic word embeddings can capture rich and interpretable semantic informat ion and uncertainty (Vilnis & McCallum, 2014; Athiwaratkun & Wilson, 2017). The uncertainty information can be particularly meaningful in capturing entailment r elationships – whereby general words such as “entity” correspond to broad distri butions that encompass more specific words such as “animal” or “instrument”. We introduce density order embeddings, which learn hierarchical representations thr ough encapsulation of probability distributions. In particular, we propose simpl e yet effective loss functions and distance metrics, as well as graph-based sche mes to select negative samples to better learn hierarchical probabilistic repres entations. Our approach provides state-of-the-art performance on the WordNet hyp ernym relationship prediction task and the challenging HyperLex lexical entailme nt dataset – while retaining a rich and interpretable probabilistic representati on.
****
Self-Supervised Learning of Object Motion Through Adversarial Video Prediction

Alex X. Lee,Frederik Ebert,Richard Zhang,Chelsea Finn,Pieter Abbeel,Sergey Levine
Can we build models that automatically learn about object motion from raw, unlabeled videos? In this paper, we study the problem of multi-step video prediction, where the goal is to predict a sequence of future frames conditioned on a short context. We focus specifically on two aspects of video prediction: accurately modeling object motion, and producing naturalistic image predictions. Our model is based on a flow-based generator network with a discriminator used to improve prediction quality. The implicit flow in the generator can be examined to determine its accuracy, and the predicted images can be evaluated for image quality. We argue that these two metrics are critical for understanding whether the model has effectively learned object motion, and propose a novel evaluation benchmark based on ground truth object flow. Our network achieves state-of-the-art results in terms of both the realism of the predicted images, as determined by human judges, and the accuracy of the predicted flow. Videos and full results can be viewed on the supplementary website: \url{https://sites.google.com/site/omvideoprediction}.
****
Do Deep Reinforcement Learning Algorithms really Learn to Navigate?
Shurjo Banerjee,Vikas Dhiman,Brent Griffin,Jason J. Corso
Deep reinforcement learning (DRL) algorithms have demonstrated progress in learning to find a goal in challenging environments. As the title of the paper by Mirowski et al. (2016) suggests, one might assume that DRL-based algorithms are able to “learn to navigate” and are thus ready to replace classical mapping and path-planning algorithms, at least in simulated environments. Yet, from experiments and analysis in this earlier work, it is not clear what strategies are used by these algorithms in navigating the mazes and finding the goal. In this paper, we pose and study this underlying question: are DRL algorithms doing some form of mapping and/or path-planning? Our experiments show that the algorithms are not memorizing the maps of mazes at the testing stage but, rather, at the training stage. Hence, the DRL algorithms fall short of qualifying as mapping or path-planning algorithms with any reasonable definition of mapping. We extend the experiments in Mirowski et al. (2016) by separating the set of training and testing maps and by a more ablative coverage of the space of experiments. Our systematic experiments show that the NavA3C-D1-D2-L algorithm, when trained and tested on the same maps, is able to choose the shorter paths to the goal. However, when tested on unseen maps the algorithm utilizes a wall-following strategy to find the goal without doing any mapping or path planning.
****
Bit-Regularized Optimization of Neural Nets
Mohamed Amer,Aswin Raghavan,Graham W. Taylor,Sek Chai
We present a novel regularization strategy for training neural networks which we call ``BitNet''. The parameters of neural networks are usually unconstrained and have a dynamic range dispersed over a real valued range. Our key idea is to control the expressive power of the network by dynamically quantizing the range and set of values that the parameters can take. We formulate this idea using a novel end-to-end approach that regularizes a typical classification loss function. Our regularizer is inspired by the Minimum Description Length (MDL) principle. For each layer of the network, our approach optimizes a translation and scaling factor along with integer-valued parameters. We empirically compare BitNet to an equivalent unregularized model on the MNIST and CIFAR-10 datasets. We show that BitNet converges faster to a superior quality solution. Additionally, the resulting model is significantly smaller in size due to the use of integer instead of floating-point parameters.
****
Semantic Code Repair using Neuro-Symbolic Transformation Networks
Jacob Devlin,Jonathan Uesato,Rishabh Singh,Pushmeet Kohli
We study the problem of semantic code repair, which can be broadly defined as automatically fixing non-syntactic bugs in source code. The majority of past work in semantic code repair assumed access to unit tests against which candidate rep

airs could be validated. In contrast, the goal here is to develop a strong statistical model to accurately predict both bug locations and exact fixes without access to information about the intended correct behavior of the program. Achieving such a goal requires a robust contextual repair model, which we train on a large corpus of real-world source code that has been augmented with synthetically injected bugs. Our framework adopts a two-stage approach where first a large set of repair candidates are generated by rule-based processors, and then these candidates are scored by a statistical model using a novel neural network architecture which we refer to as Share, Specialize, and Compete. Specifically, the architecture (1) generates a shared encoding of the source code using an RNN over the abstract syntax tree, (2) scores each candidate repair using specialized network modules, and (3) then normalizes these scores together so they can compete against one another in comparable probability space. We evaluate our model on a real-world test set gathered from GitHub containing four common categories of bugs. Our model is able to predict the exact correct repair 41% of the time with a single guess, compared to 13% accuracy for an attentional sequence-to-sequence model.
****

MaskGAN: Better Text Generation via Filling in the _______
William Fedus,Ian Goodfellow,Andrew M. Dai
Neural text generation models are often autoregressive language models or seq2seq models. Neural autoregressive and seq2seq models that generate text by sampling words sequentially, with each word conditioned on the previous model, are state-of-the-art for several machine translation and summarization benchmarks. These benchmarks are often defined by validation perplexity even though this is not a direct measure of sample quality. Language models are typically trained via maximum likelihood and most often with teacher forcing. Teacher forcing is well-suited to optimizing perplexity but can result in poor sample quality because generating text requires conditioning on sequences of words that were never observed at training time. We propose to improve sample quality using Generative Adversarial Network (GANs), which explicitly train the generator to produce high quality samples and have shown a lot of success in image generation. GANs were originally to designed to output differentiable values, so discrete language generation is challenging for them. We introduce an actor-critic conditional GAN that fills in missing text conditioned on the surrounding context. We show qualitatively and quantitatively, evidence that this produces more realistic text samples compared to a maximum likelihood trained model.
****

Boundary Seeking GANs
R Devon Hjelm,Athul Paul Jacob,Adam Trischler,Gerry Che,Kyunghyun Cho,Yoshua Bengio
Generative adversarial networks are a learning framework that rely on training a discriminator to estimate a measure of difference between a target and generated distributions. GANs, as normally formulated, rely on the generated samples being completely differentiable w.r.t. the generative parameters, and thus do not work for discrete data. We introduce a method for training GANs with discrete data that uses the estimated difference measure from the discriminator to compute importance weights for generated samples, thus providing a policy gradient for training the generator. The importance weights have a strong connection to the decision boundary of the discriminator, and we call our method boundary-seeking GANs (BGANs). We demonstrate the effectiveness of the proposed algorithm with discrete image and character-based natural language generation. In addition, the boundary-seeking objective extends to continuous data, which can be used to improve stability of training, and we demonstrate this on Celeba, Large-scale Scene Understanding (LSUN) bedrooms, and Imagenet without conditioning.
****

Active Learning for Convolutional Neural Networks: A Core-Set Approach
Ozan Sener,Silvio Savarese
Convolutional neural networks (CNNs) have been successfully applied to many recognition and learning tasks using a universal recipe; training a deep model on a

very large dataset of supervised examples. However, this approach is rather restrictive in practice since collecting a large set of labeled images is very expensive. One way to ease this problem is coming up with smart ways for choosing images to be labelled from a very large collection (i.e. active learning).

Our empirical study suggests that many of the active learning heuristics in the literature are not effective when applied to CNNs when applied in batch setting. Inspired by these limitations, we define the problem of active learning as core-set selection, i.e. choosing set of points such that a model learned over the selected subset is competitive for the remaining data points. We further present a theoretical result characterizing the performance of any selected subset using the geometry of the datapoints. As an active learning algorithm, we choose the subset which is expected to yield best result according to our characterization. Our experiments show that the proposed method significantly outperforms existing approaches in image classification experiments by a large margin.

****

A Deep Reinforced Model for Abstractive Summarization

Romain Paulus,Caiming Xiong,Richard Socher

Attentional, RNN-based encoder-decoder models for abstractive summarization have achieved good performance on short input and output sequences. For longer documents and summaries however these models often include repetitive and incoherent phrases. We introduce a neural network model with a novel intra-attention that attends over the input and continuously generated output separately, and a new training method that combines standard supervised word prediction and reinforcement learning (RL).

Models trained only with supervised learning often exhibit "exposure bias" - they assume ground truth is provided at each step during training.

However, when standard word prediction is combined with the global sequence prediction training of RL the resulting summaries become more readable.

We evaluate this model on the CNN/Daily Mail and New York Times datasets. Our model obtains a 41.16 ROUGE-1 score on the CNN/Daily Mail dataset, an improvement over previous state-of-the-art models. Human evaluation also shows that our model produces higher quality summaries.

****

Semantic Interpolation in Implicit Models

Yannic Kilcher,Aurelien Lucchi,Thomas Hofmann

In implicit models, one often interpolates between sampled points in latent space. As we show in this paper, care needs to be taken to match-up the distributional assumptions on code vectors with the geometry of the interpolating paths. Otherwise, typical assumptions about the quality and semantics of in-between points may not be justified. Based on our analysis we propose to modify the prior code distribution to put significantly more probability mass closer to the origin. As a result, linear interpolation paths are not only shortest paths, but they are also guaranteed to pass through high-density regions, irrespective of the dimensionality of the latent space. Experiments on standard benchmark image datasets demonstrate clear visual improvements in the quality of the generated samples and exhibit more meaningful interpolation paths.

****

Active Neural Localization

Devendra Singh Chaplot,Emilio Parisotto,Ruslan Salakhutdinov

Localization is the problem of estimating the location of an autonomous agent from an observation and a map of the environment. Traditional methods of localization, which filter the belief based on the observations, are sub-optimal in the number of steps required, as they do not decide the actions taken by the agent. We propose "Active Neural Localizer", a fully differentiable neural network that learns to localize efficiently. The proposed model incorporates ideas of traditional filtering-based localization methods, by using a structured belief of the state with multiplicative interactions to propagate belief, and combines it with a policy model to minimize the number of steps required for localization. Active

 Neural Localizer is trained end-to-end with reinforcement learning. We use a va
riety of simulation environments for our experiments which include random 2D maz
es, random mazes in the Doom game engine and a photo-realistic environment in th
e Unreal game engine. The results on the 2D environments show the effectiveness
of the learned policy in an idealistic setting while results on the 3D environme
nts demonstrate the model's capability of learning the policy and perceptual mod
el jointly from raw-pixel based RGB observations. We also show that a model trai
ned on random textures in the Doom environment generalizes well to a photo-reali
stic office space environment in the Unreal engine.
****
Ego-CNN: An Ego Network-based Representation of Graphs Detecting Critical Struct
ures
Ruo-Chun Tzeng,Shan-Hung Wu
While existing graph embedding models can generate useful embedding vectors that
 perform well on graph-related tasks, what valuable information can be jointly l
earned by a graph embedding model is less discussed. In this paper, we consider
the possibility of detecting critical structures by a graph embedding model. We
propose Ego-CNN to embed graph, which works in a local-to-global manner to take
advantages of CNNs that gradually expanding the detectable local regions on the
graph as the network depth increases. Critical structures can be detected if Ego
-CNN is combined with a supervised task model. We show that Ego-CNN is (1) compe
titive to state-of-the-art graph embeddings models, (2) can nicely work with CNN
s visualization techniques to show the detected structures, and (3) is efficient
 and can incorporate with scale-free priors, which commonly occurs in social net
work datasets, to further improve the training efficiency.
****
GradNorm: Gradient Normalization for Adaptive Loss Balancing in Deep Multitask N
etworks
Zhao Chen,Vijay Badrinarayanan,Chen-Yu Lee,Andrew Rabinovich
Deep multitask networks, in which one neural network produces multiple predictiv
e outputs, are more scalable and often better regularized than their single-task
 counterparts. Such advantages can potentially lead to gains in both speed and p
erformance, but multitask networks are also difficult to train without finding t
he right balance between tasks. We present a novel gradient normalization (GradN
orm) technique which automatically balances the multitask loss function by direc
tly tuning the gradients to equalize task training rates. We show that for vario
us network architectures, for both regression and classification tasks, and on b
oth synthetic and real datasets, GradNorm improves accuracy and reduces overfitt
ing over single networks, static baselines, and other adaptive multitask loss ba
lancing techniques. GradNorm also matches or surpasses the performance of exhaus
tive grid search methods, despite only involving a single asymmetry hyperparamet
er α. Thus, what was once a tedious search process which incurred exponen
tially more compute for each task added can now be accomplished within a few tra
ining runs, irrespective of the number of tasks. Ultimately, we hope to demonstr
ate that gradient manipulation affords us great control over the training dynami
cs of multitask networks and may be one of the keys to unlocking the potential o
f multitask learning.
****
Improving image generative models with human interactions
Andrew Kyle Lampinen,David So,Douglas Eck,Fred Bertsch
GANs provide a framework for training generative models which mimic a data distr
ibution. However, in many cases we wish to train a generative model to optimize
some auxiliary objective function within the data it generates, such as making m
ore aesthetically pleasing images. In some cases, these objective functions are
difficult to evaluate, e.g. they may require human interaction. Here, we develop
 a system for efficiently training a GAN to increase a generic rate of positive
user interactions, for example aesthetic ratings. To do this, we build a model o
f human behavior in the targeted domain from a relatively small set of interacti
ons, and then use this behavioral model as an auxiliary loss function to improve
 the generative model. As a proof of concept, we demonstrate that this system is

successful at improving positive interaction rates simulated from a variety of objectives, and characterize s
****

EXPLORING NEURAL ARCHITECTURE SEARCH FOR LANGUAGE TASKS
Minh-Thang Luong,David Dohan,Adams Wei Yu,Quoc V. Le,Barret Zoph,Vijay Vasudevan
Neural architecture search (NAS), the task of finding neural architectures automatically, has recently emerged as a promising approach for unveiling better models over human-designed ones. However, most success stories are for vision tasks and have been quite limited for text, except for a small language modeling setup. In this paper, we explore NAS for text sequences at scale, by first focusing on the task of language translation and later extending to reading comprehension. From a standard sequence-to-sequence models for translation, we conduct extensive searches over the recurrent cells and attention similarity functions across two translation tasks, IWSLT English-Vietnamese and WMT German-English. We report challenges in performing cell searches as well as demonstrate initial success on attention searches with translation improvements over strong baselines. In addition, we show that results on attention searches are transferable to reading comprehension on the SQuAD dataset.
****

Feature Map Variational Auto-Encoders
Lars Maaløe,Ole Winther
There have been multiple attempts with variational auto-encoders (VAE) to learn powerful global representations of complex data using a combination of latent stochastic variables and an autoregressive model over the dimensions of the data. However, for the most challenging natural image tasks the purely autoregressive model with stochastic variables still outperform the combined stochastic autoregressive models. In this paper, we present simple additions to the VAE framework that generalize to natural images by embedding spatial information in the stochastic layers. We significantly improve the state-of-the-art results on MNIST, OMNIGLOT, CIFAR10 and ImageNet when the feature map parameterization of the stochastic variables are combined with the autoregressive PixelCNN approach. Interestingly, we also observe close to state-of-the-art results without the autoregressive part. This opens the possibility for high quality image generation with only one forward-pass.

****

Critical Percolation as a Framework to Analyze the Training of Deep Networks
Zohar Ringel,Rodrigo Andrade de Bem
In this paper we approach two relevant deep learning topics: i) tackling of graph structured input data and ii) a better understanding and analysis of deep networks and related learning algorithms. With this in mind we focus on the topological classification of reachability in a particular subset of planar graphs (Mazes). Doing so, we are able to model the topology of data while staying in Euclidean space, thus allowing its processing with standard CNN architectures. We suggest a suitable architecture for this problem and show that it can express a perfect solution to the classification task. The shape of the cost function around this solution is also derived and, remarkably, does not depend on the size of the maze in the large maze limit. Responsible for this behavior are rare events in the dataset which strongly regulate the shape of the cost function near this global minimum. We further identify an obstacle to learning in the form of poorly performing local minima in which the network chooses to ignore some of the inputs. We further support our claims with training experiments and numerical analysis of the cost function on networks with up to 128 layers.
****