Conference Report - NIPS 2017

Written by: Tassilo Klein, Marius Lehne, Brian Clarke, Steven Jaeger

NIPS, Long Beach, USA, December 4th - 9th

- General
- Selection of Interesting Papers
- Conclusion

General

The *Conference on Neural Information Processing Systems (NIPS)* took place between December 4th and 9th in Long Beach, CA, USA.

As one of the top machine learning and computational neuroscience conferences, this year's NIPS was a complete success experiencing a larger rush of attendees compared to previous years. Apart from the rising number of participants, the conference has also seen a strong increase in submitted papers. Of the total 3240 papers submitted, 679 papers were accepted resulting in a 21% acceptance rate compared to last year's 24%. This is an indicator of the conference remaining highly competitive despite its overall growing popularity.

As an official sponsor of NIPS, the SAP team, including ML researchers and data scientists, were on site to present SAP's machine learning solutions and research projects. Apart from connecting with the broader academic research community during the plethora of talks, workshops and tutorials, the ML research team presented their recent work on Federated Learning with Differential Privacy in the workshop for Machine Learning on the Phone and other Consumer Devices. Furthermore, our research partners from University of Amsterdam presented their work Emergence of Language with Multi-agent Games: Learning to Communicate with Sequences of Symbols.

Topics and Trends

As deep learning has developed to become a de-facto standard commodity in data science and research with numerous tools now readily available for non-ML experts, many new sub-domains of deep learning have begun to spring up. While the pre-deep learning area was dominated by feature engineering, the ML community has been experiencing a network architecture engineering era afterwards. However, now there is a massive push to establish theoretical foundations for all the methods proposed over the last years, which goes beyond proposing new (deeper) architectures. Besides the standard machine learning topics and trends therein, social topics in ML have started to be part of the conference agenda. This can be attributed to the emergence and integration of machine learning services into our daily lives including all the pros . As a result, there were a lot of talks on ethics, bias, privacy as well as a workshop addressing ML for the developing world.

- **Generative Adversarial Networks** Similar to other conferences, the topic of generative models has seen quite some interest at NIPS, where particularly Generative Adversarial Networks (GANs) made up a significant amount of the contributions. However, GANs remain difficult to train. First, we could observe rather theoretical work determining and

analyzing reasons for the instability of the learning process. Then, there were numerous modifications of GANs to further improve this process. Additionally, we saw many contributions that used GANs for various types of applications (e.g. semi-supervised learning).

- **Reinforcement Learning** Reinforcement Learning (RL) has seen a lot of progress for many applications, as well as various research trends. However, RL still requires huge amounts of data and a very long time to learn, while at the same time suffering from poor generalization. In order to alleviate this problem much research is conducted in the domain of meta reinforcement learning approaches that can adapt quickly to new environments. An interesting paper in this domain is Learning to reinforcement learn, as well as Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. Similarly, the notion of Hierarchical Reinforcement Learning tries to breakup tasks into smaller reoccurring manageable components, each controlled by sub-policies instead of one monolithic policy that might be harder to train. One paper trying to combine the idea of Meta RL and Hierarchical RL is Meta Learning Shared Hierarchies. The topic of RL and especially Hierarchical RL was also prominent in the workshop on conversational applications. The basic idea is to have several domain specific agents that are orchestrated by a manager agent. Another interesting extension was presented in Neural Map: Structured Memory for Deep Reinforcement Learning where the authors extend the standard memory architectures (usually an RNN) with dynamic memory networks. The topic of interpretability and 'white boxing' has gained traction in RL. An intriguing application can be seen in Natural Language Policy Search. In her talk, Joelle Pineau provided somewhat shocking insights on reproducibility: The difference in outcomes for two Reinforcement Learning agents were statistically significant depending on the random seed of the initialisation.

- **Bayesian Deep Learning and Deep Bayesian Learning** Naturally, Deep Learning suffers from several shortcomings that limit wider applicability such as lack of interpretability, notion of confidence in prediction, as well as missing mathematical foundation. Bayesian Deep Learning tries to compensate these issues by means of combining Deep Learning with Bayesian probability theory. For more details, check out Uncertainty in Deep Learning, as well as the white paper Deep Learning: A Bayesian Perspective.

- **Meta-Learning** In meta-learning, which is often also referred to as "learning to learn", the deep learning model is itself a learning algorithm. The notion behind it being that when (meta-)training such a model, one will be able to learn a procedure that can learn in a more efficient manner through better generalization. Exemplary papers in this field are Learning to learn by gradient descent by gradient descent and Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. Not surprisingly, meta-learning is a promising approach for few-shot learning, which seeks to learn new concepts from just a handful of training examples.

- **Theoretical Foundations of Deep Learning** Modern deep learning architectures have enormous capacity, which theoretically allows them to easily memorize the data sets they were trained on. Yet, the models deliver state of the art results. Thus, one important topic, which up to now remains poorly understood, is to find explanations for the generalization performance of deep neural networks. More details can be found in the work Understanding Deep Learning requires rethink generalization presented at [ICML] (https://icml.cc/Conferences/2017) this year. The lack of theoretical foundations in deep learning has also been criticized by Ali Rahimi in his acceptance speech for the test of time award. As within other conferences in the field, there were many papers focussing on the various efforts to move towards a better understanding of the generalization performance of deep neural networks.

- **Fairness in Machine Learning** Fairness in machine learning is a research subfield concerned with the notion of avoiding unintended discrimination. This topic has been actively debated in various media reports in recent times. The concern arises from the fact that there may exist (implicit) bias in training data or in the formulation of algorithms, which may negatively impact society. As the research community has realized the need to address these concerns, the topic has received great attention throughout the whole conference. In their excellent tutorial, Moritz Hardt and Solon Barocas brought together the legal and the technical perspective of fairness in machine learning. This was followed up by an Invited Talk of Kate Crawford, in which she spoke of the impact of unfair machine learning methods on society. She pointed out that fairness is not only a technical problem but also a social one. Different papers were presented during the conference proposing methods (e.g. counterfactual machine learning) to address the different sources of bias.

- **Privacy in Machine Learning** In many scenarios, practitioners and scientists have to work on sensitive data. Thus, preserving privacy in a machine learning context has become an essential topic for numerous tasks. This was also well reflected in the conference's contributions, as many researchers proposed modifications to introduce privacy measures into various established machine learning methods. Differential privacy has become the method of choice for measuring privacy risks. If you are interested in a paper with the elementary definitions and formalisms of this topic, please check out Deep Learning with Differential Privacy.

- **Explainable and Interpretable Machine Learning** Many currently successful deep learning methods are black-box methods. Even though practitioners and domain experts have identified that explaining a model's outcome is highly important, there is an ongoing discourse about the necessity of explainable machine learning in practice. Yet, opening this black box through explainable and interpretable machine learning was well represented throughout the conference in various formats and it was noticeable that different approaches were presented or improved. As many proposed methods fail to live up to expectations, there is also a discussion about short comings and limitations. Due to this, there is a move towards a better definition of what such methods should achieve and how they can be evaluated in a better way.

- **Learning on Graphs and Manifolds** Various lines of work, summarized in an interesting tutorial, in recent years have pushed towards extending the success of convolutional neural networks to graphs. Graphs are generalizations of the regular lattices found in text (1D grids) or images (3D grids) and are relevant to many domains e.g. social networks, molecules, and 3D shapes, which become graphs when represented as a discrete mesh. There are multiple possible generalizations of CNNs to graphs, with the two main paradigms being based on spectral and spatial interpretations of CNNs. At NIPS, the advantages and disadvantages of the two paradigms were presented, along with a number of intriguing applications ranging from recommender systems to quantum chemistry and particle physics.

Selection of Interesting Papers

Due to the high number of papers at NIPS (over 1000 papers), we can only present a handful of papers that caught our attention. However, there were a lot more interesting papers presented and we therefore strongly recommend having a look at the list of papers here, as well as the workshop sites.

For illustrative purposes, selected figures were taken from the papers.

- **Emergence of Language with Multi-agent Games: Learning to Communicate with Sequences of Symbols (*SAP ML Research Sponsored*)**

Abstract

"Learning to communicate through interaction, rather than relying on explicit supervision, is often considered a prerequisite for developing a general AI. We study a setting where two agents engage in playing a referential game and, from scratch, develop a communication protocol necessary to succeed in this game. Unlike previous work, we require that messages they exchange, both at train and test time, are in the form of a language (i.e. sequences of discrete symbols). We compare a reinforcement learning approach and one using a differentiable relaxation (straight-through Gumbel-softmax estimator) and observe that the latter is much faster to converge and it results in more effective protocols. Interestingly, we also observe that the protocol we induce by optimizing the communication success exhibits a degree of compositionality and variability (i.e. the same information can be phrased in different ways), both properties characteristic of natural languages. As the ultimate goal is to ensure that communication is accomplished in natural language, we also perform experiments where we inject prior information about natural language into our model and study properties of the resulting protocol."

> **Why we think it's interesting:** It is a quite intriguing setting to have agents engaging in playing a referential game and, basically from scratch, develop a communication protocol (language) necessary to succeed in this game. Particularly interesting is the notion that the developed language implements some kind of hierarchical structure, i.e. the word order matters in the code.

- **Differentially Private Federated Learning: A Client Level Perspective (*SAP ML Research contribution*)**)

Abstract

"Federated learning is a recent advance in privacy protection. In this context, a trusted curator aggregates parameters optimized in decentralized fashion by multiple clients. The resulting model is then distributed back to all clients, ultimately converging to a joint representative model without explicitly having to share the data. However, the protocol is vulnerable to differential attacks, which could originate from any party contributing during federated optimization. In such an attack, a client's contribution during training and information about their data set is revealed through analyzing the distributed model.

We tackle this problem and propose an algorithm for client sided differential privacy preserving federated optimization. The aim is to hide clients' contributions during training, balancing the trade-off between privacy loss and model performance. Empirical studies suggest that given a sufficiently large number of participating clients, our proposed procedure can maintain client-level differential privacy at only a minor cost in model performance."

Why we think it's interesting: For many scenarios and applications the number of participants learning a model can be limited. This approach shows that it is possible to train machine learning models under privacy constraints even in settings where the number of participant clients is not extremely large, with a minor decrease in accuracy. Our findings are of particular interest for hospitals, companies or any kind of institution that wants to benefit from generalized prediction models but is bound to strong privacy requirements. Federated learning with differential privacy enables them to benefit from a generalized model learned by many peer contributors without the need of centralizing data or taking the risk of exposing private information.

- **Gradient descent GAN optimization is locally stable**

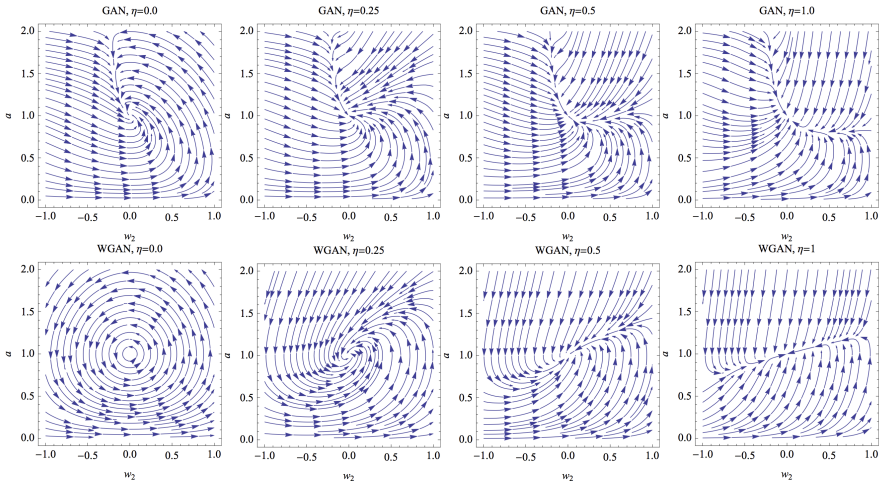


Figure 3: Streamline plots around the equilibrium $(0, 1)$ for the conventional GAN (top) and the WGAN (bottom) for $\eta = 0$ (vanilla updates) and $\eta = 0.25, 0.5, 1$ (left to right).

Abstract

"Despite the growing prominence of generative adversarial networks (GANs), optimization in GANs is still a poorly understood topic. In this paper, we analyze the"gradient descent" form of GAN optimization i.e., the natural setting where we simultaneously take small gradient steps in both generator and discriminator parameters. We show that even though GAN optimization does not correspond to a convex-concave game (even for simple parameterizations), under proper conditions, equilibrium points of this optimization procedure are still locally asymptotically stable for the traditional GAN formulation. On the other hand, we show that the recently proposed Wasserstein GAN can have non-convergent limit cycles near equilibrium. Motivated by this stability analysis, we propose an additional regularization term for gradient descent GAN updates, which is able to guarantee local stability for both the WGAN and the traditional GAN, and also shows practical promise in speeding up convergence and addressing mode collapse."

> **Why we think it's interesting:** The paper studies the stability properties of GANs by means of tools from non-linear systems theory. Interestingly enough, standard GANs seem to be locally asymptotically stable. However, this is not the case for the recently proposed popular variant Wasserstein GAN, which typically is less subject to issues such as mode collapse.

- **Variance-based Regularization with Convex Objectives** (*Best Paper Award*)

Abstract

"We develop an approach to risk minimization and stochastic optimization that provides a convex surrogate for variance, allowing near-optimal and computationally efficient trading between approximation and estimation error. Our approach builds off of techniques for distributionally robust optimization and Owen's empirical likelihood, and we provide a number of finite-sample and asymptotic results characterizing the theoretical performance of the estimator. In particular, we show that our procedure comes with certificates of optimality, achieving (in some scenarios) faster rates of convergence than empirical risk minimization by virtue of automatically balancing bias and variance. We give corroborating empirical evidence showing that in practice, the estimator indeed trades between variance and absolute performance on a training sample, improving out-of-sample (test) performance over standard empirical risk minimization for a number of classification problems."

Why we think it's interesting: Learning algorithms are subject to bias (approximation error) and variance (estimation) error, which is commonly known as the bias-variance dilemma. Basically, this prevents supervised learning algorithms from generalizing beyond the training set. Conventional neural networks are optimized by what is referred to as Empirical Risk Minimization (ERM), which is the minimization of the mean empirical error on training data. The approach proposed in the paper represents a trade-off between bias and variance components. This is achieved by using a worst-case loss, which typically performs better than the average as in ERM. Implicitly, the algorithm increases the weight of hard (or rare) examples in the training set, instead of weighting them equally (ERM). Therefore, minimization with respect to this robust error bound can compensate unequal class distributions (rare classes). This is posed as a convex-optimization problem and can thus be solved efficiently.

- **Net-Trim: Convex Pruning of Deep Neural Networks with Performance Guarantee**

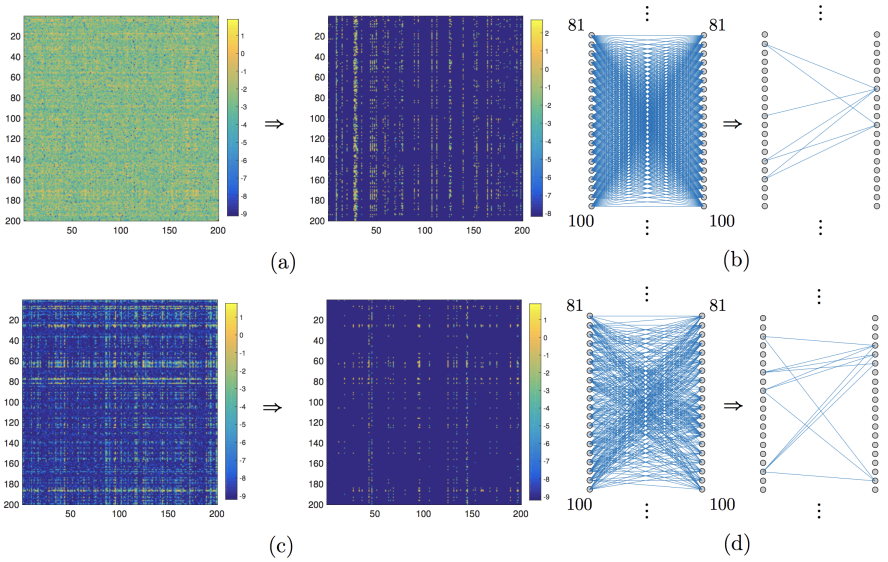


Figure 1: Net-Trim pruning performance on classification of points within nested spirals; (a) left: the weighted adjacency matrix relating the two hidden layers after training; right: the adjacency matrix after the application of Net-Trim causing more than 93% of the weights to vanish; (b) partial network topology relating neurons 81 to 100 of the hidden layers, before and after retraining; (c) left: the adjacency matrix after training the network with dropout and ℓ_1 regularization; right: Net-Trim is yet able to find a model which is over 7 times sparser than the model on the left; (d) partial network topology before and after retraining for panel (c)

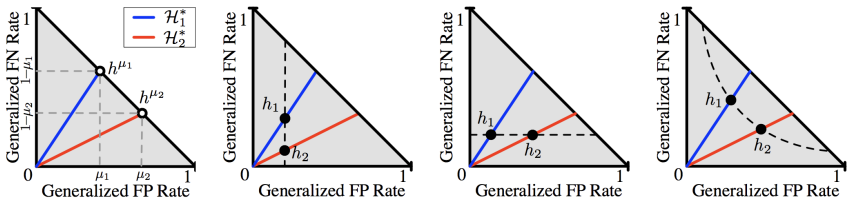
Abstract

"We introduce and analyze a new technique for model reduction for deep neural networks. While large networks are theoretically capable of learning arbitrarily complex models, overfitting and model redundancy negatively affects the prediction accuracy and model variance. Our Net-Trim algorithm prunes (sparsifies) a trained network layer-wise, removing connections at each layer by solving a convex optimization program. This program seeks a sparse set of weights at each layer that keeps the layer inputs and outputs consistent with the originally trained model. The algorithms and associated analysis are applicable to neural networks operating with the rectified linear unit (ReLU) as the nonlinear activation. We present both parallel and cascade versions of the algorithm. While the latter can achieve slightly simpler models with the same generalization performance, the former can be computed in a distributed manner. In both cases, Net-Trim significantly reduces the number of connections in the network, while also providing enough regularization to slightly reduce the generalization error.

We also provide a mathematical analysis of the consistency between the initial network and the retrained model. To analyze the model sample complexity, we derive the general sufficient conditions for the recovery of a sparse transform matrix. For a single layer taking independent Gaussian random vectors of length N as inputs, we show that if the network response can be described using a maximum number of s non-zero weights per node, these weights can be learned from O(slogN) samples."

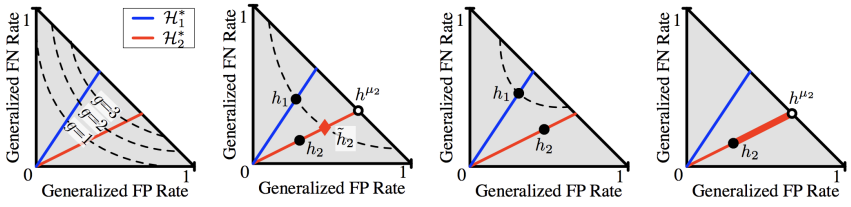
Why we think it's interesting: There has been a trend to make Neural Network architecture more and more complex. However, this leads to longer training and inference times, as well as larger memory footprint. Therefore, the topic of efficient deep learning has enjoyed more and more attention recently. One possible approach is the so-called model compression. This paper follows a similar notion by means of drastically sparsifying connections within neural network, without compromising the accuracy. This is achieved by applying a generic layer-wise post-processing efficient convex-programming routine.

- **On Fairness and Calibration**



(a) Possible cal. classifiers $\mathcal{H}_1^*, \mathcal{H}_2^*$ (blue/red). (b) Satisfying cal. and equal F.P. rates. (c) Satisfying cal. and equal F.N. rates. (d) Satisfying cal. and a general constraint.

Figure 1: Calibration, trivial classifiers, and equal-cost constraints – plotted in the false-pos./false-neg. plane. $\mathcal{H}_1^*, \mathcal{H}_2^*$ are the set of cal. classifiers for the two groups, and h^{μ_1}, h^{μ_2} are trivial classifiers.



(a) Level-order curves of cost. Low cost implies low error rates. (b) Usually, there is a calibrated classifier $\tilde{h}_2$ with the same cost of h_1. (c) Cal. and equal-cost are incompatible if h_1 has high error. (d) Possible cal. classifiers for G_2 (bold red) by mixing h_2 and h^{μ_2}.

Figure 2: Calibration-Preserving Parity through interpolation.

Abstract

"The machine learning community has become increasingly concerned with the potential for bias and discrimination in predictive models. This has motivated a growing line of work on what it means for a classification procedure to be"fair." In this paper, we investigate the tension between minimizing error disparity across different population groups while maintaining calibrated probability estimates. We show that calibration is compatible only with a single error constraint (i.e. equal false-negatives rates across groups), and show that any algorithm that satisfies this relaxation is no better than randomizing a percentage of predictions for an existing classifier. These unsettling findings, which extend and generalize existing results, are empirically confirmed on several datasets."

Why we think it's interesting: Fairness tries to minimize the effect of factors such as race/gender/sexual orientation in data on the decision outcome. Beside fairness, it is desirable that machine learning algorithms produce predictions that reflect reality, i.e. that the predicted probability estimates represent the true correctness likelihood. This coincides with the notion of calibration. However, often the probabilities produced by softmax do not represent reality. There are approaches that seek to modify the probabilities by means of calibration, as discussed in an interesting paper submitted to ICML'18 On Calibration of Modern Neural Networks. However, as it turns out calibration clashes with the notion of fairness/unbiasedness when different population groups are considered.

- **Deep Sets**

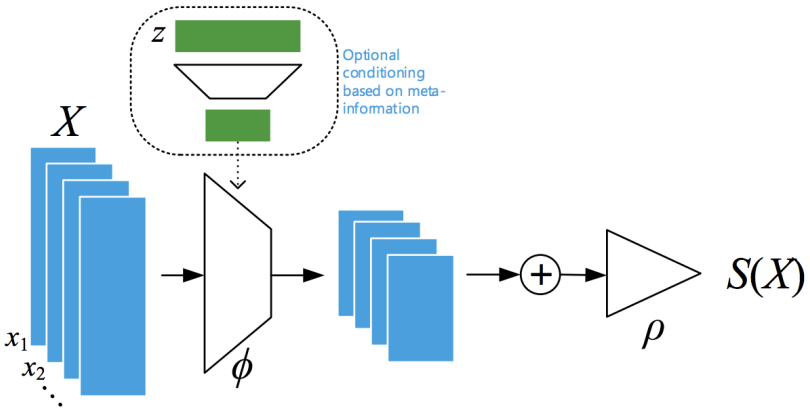


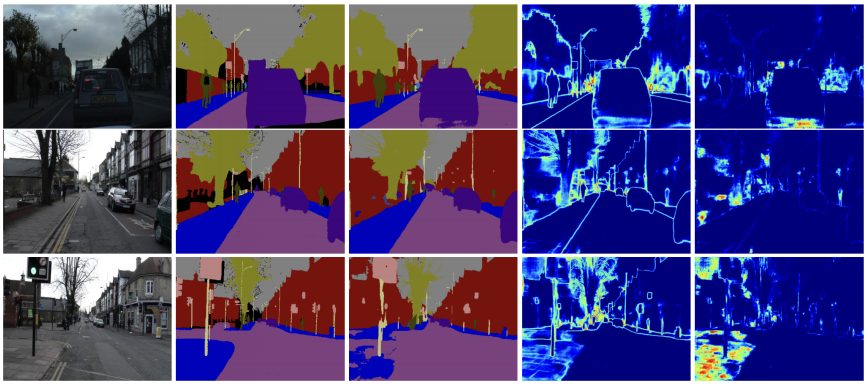
Figure 2. Architecture of deep sets

Abstract

"In this paper, we study the problem of designing objective functions for machine learning problems defined on finite sets. In contrast to traditional objective functions defined for machine learning problems operating on finite dimensional vectors, the new objective functions we propose are operating on finite sets and are invariant to permutations. Such problems are widespread, ranging from estimation of population statistics, via anomaly detection in piezometer data of embankment dams, to cosmology. Our main theorem characterizes the permutation invariant objective functions and provides a family of functions to which any permutation invariant objective function must belong. This family of functions has a special structure which enables us to design a deep network architecture that can operate on sets and which can be deployed on a variety of scenarios including both unsupervised and supervised learning tasks. We demonstrate the applicability of our method on population statistic estimation, point cloud classification, set expansion, and image tagging."

Why we think it's interesting: For many applications the input is order invariant and may vary in size. These characteristics can be formalized as sets. Conventional neural networks, however, require fixed representations and are not order-invariant. This paper formalized the functional aspects for set representation and proposes a deep network that can deal with inputs that have set character.

- **What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?**



(a) Input Image (b) Ground Truth (c) Semantic Segmentation (d) Aleatoric Uncertainty (e) Epistemic Uncertainty

Figure 1: **Illustrating the difference between aleatoric and epistemic uncertainty** for semantic segmentation on the CamVid dataset [8]. *Aleatoric* uncertainty captures noise inherent in the observations. In (d) our model exhibits increased aleatoric uncertainty on object boundaries and for objects far from the camera. *Epistemic* uncertainty accounts for our ignorance about which model generated our collected data. This is a notably different measure of uncertainty and in (e) our model exhibits increased epistemic uncertainty for semantically and visually challenging pixels. The bottom row shows a failure case of the segmentation model when the model fails to segment the footpath due to increased epistemic uncertainty, but not aleatoric uncertainty.

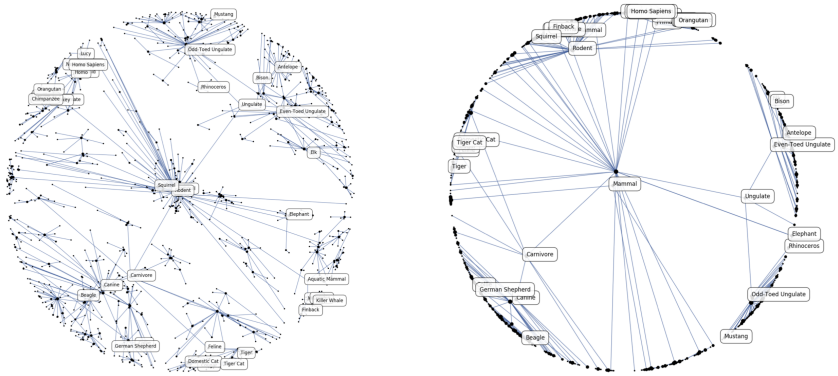
Abstract

"There are two major types of uncertainty one can model. Aleatoric uncertainty captures noise inherent in the observations. On the other hand, epistemic uncertainty accounts for uncertainty in the model – uncertainty which can be explained away given enough data. Traditionally it has been difficult to model epistemic uncertainty in computer vision, but with new Bayesian deep learning tools this is now possible. We study the benefits of modeling epistemic vs. aleatoric uncertainty in Bayesian deep learning models for vision tasks. For this we present a Bayesian deep learning framework combining input-dependent aleatoric uncertainty together with epistemic uncertainty. We study models under the framework with per-pixel semantic segmentation and depth regression tasks. Further, our explicit uncertainty formulation leads to new loss functions for these tasks, which can be interpreted as learned attenuation. This makes the loss more robust to noisy data, also giving new state-of-the-art results on segmentation and depth regression benchmarks."

Why we think it's interesting: Uncertainty modeling is a a hot topic in machine learning, particularly in the context of determining false over-confidence in predictions. For those who

are interested in more details, I recommend checking out Yarin Gal's the phD-thesis. This paper studies different sources of uncertainty in the context of computer vision using Bayesian mechanisms (Bayesian Deep Learning).

- **Poincaré Embeddings for Learning Hierarchical Representations**



(a) Intermediate embedding after 20 epochs

(b) Embedding after convergence

Figure 2: Two-dimensional Poincaré embeddings of transitive closure of the WordNet mammals subtree. Ground-truth `is-a` relations of the original WordNet tree are indicated via blue edges. A Poincaré embedding with $d = 5$ achieves mean rank 1.26 and MAP 0.927 on this subtree.

Abstract

"Representation learning has become an invaluable approach for learning from symbolic data such as text and graphs. However, while complex symbolic datasets often exhibit a latent hierarchical structure, state-of-the-art methods typically learn embeddings in Euclidean vector spaces, which do not account for this property. For this purpose, we introduce a new approach for learning hierarchical representations of symbolic data by embedding them into hyperbolic space – or more precisely into an n-dimensional Poincaré ball. Due to the underlying hyperbolic geometry, this allows us to learn parsimonious representations of symbolic data by simultaneously capturing hierarchy and similarity. We introduce an efficient algorithm to learn the embeddings based on Riemannian optimization and show experimentally that Poincaré embeddings outperform Euclidean embeddings significantly on data with latent hierarchies, both in terms of representation capacity and in terms of generalization ability.

Why we think it's interesting: Manifold embeddings of concepts typically project items on a semantic level only. For word embeddings such as the famous Word2vec, this implies that words with similar semantic meaning are encoded such that their (Euclidean) distance should be minimal, i.e. their projections on the manifold are in proximity. However, this totally disregards the taxonomy of the semantic. The proposed approach tries to tackle that issue by way of e.g. defining a projection that incorporates the notion of natural taxonomy. Optimization of such an embedding is rather straightforward and simply requires a projection of the gradients.

- **Safe and Nested Subgame Solving for Imperfect-Information Games**(*Best Paper Award*)

Abstract

"In imperfect-information games, the optimal strategy in a subgame may depend on the strategy in other, unreached subgames. Thus a subgame cannot be solved in isolation and must instead consider the strategy for the entire game as a whole, unlike perfect-information games. Nevertheless, it is possible to first approximate a solution for the whole game and then improve it in individual subgames. This is referred to as subgame solving. We introduce subgame-solving techniques that outperform prior methods both in theory and practice. We also show how to adapt them, and past subgame-solving techniques, to respond to opponent actions that are outside the original action abstraction; this significantly outperforms the prior state-of-the-art approach, action translation. Finally, we show that subgame solving can be repeated as the game progresses down the game tree, leading to far lower exploitability. These techniques were a key component of Libratus, the first AI to defeat top humans in heads-up no-limit Texas hold'em poker."

Why we think it's interesting: An imperfect information game is a concept from game theory and describes games in which one participant has no information about actions that another player has taken. In practice, this type of game can be found in a wide range of applications such as popular card games, contract negotiations, price competitions and more. Through its generality the proposed techniques have broad applicability

in a wide range of those scenarios. Not only have the authors shown improvements on theoretical guarantees compared to prior techniques, but the AI which contained methods presented in this paper have stirred quite some attention by the media after beating one of the best human players in Texas Hold'em poker.

- **The marginal value of adaptive gradient methods**

Abstract

"Adaptive optimization methods, which perform local optimization with a metric constructed from the history of iterates, are becoming increasingly popular for training deep neural networks. Examples include AdaGrad, RMSProp, and Adam. We show that for simple over-parameterized problems, adaptive methods often find drastically different solutions than gradient descent (GD) or stochastic gradient descent (SGD). We construct an illustrative binary classification problem where the data is linearly separable, GD and SGD achieve zero test error, and AdaGrad, Adam, and RMSProp attain test errors arbitrarily close to half. We additionally study the empirical generalization capability of adaptive methods on several state-of-the-art deep learning models. We observe that the solutions found by adaptive methods generalize worse (often significantly worse) than SGD, even when these solutions have better training performance. These results suggest that practitioners should reconsider the use of adaptive methods to train neural networks."

Why we think it's interesting: Adaptive gradient methods and especially ADAM have become the de-facto standard for optimization algorithms in deep learning. This paper is especially interesting because the authors present empirical findings that contradict with the well-established best practices. The work further shows that the choice of the optimization technique has an effect on the overall generalization performance. Thus, it is pointed out by the authors that this effect is still not well understood. Even though the conclusions are only supported by empirical evidence, the contribution of this paper provides important insights for selecting an appropriate optimization algorithm.

- **Train longer, generalize better: Closing the generalization gap in large batch training of NN**

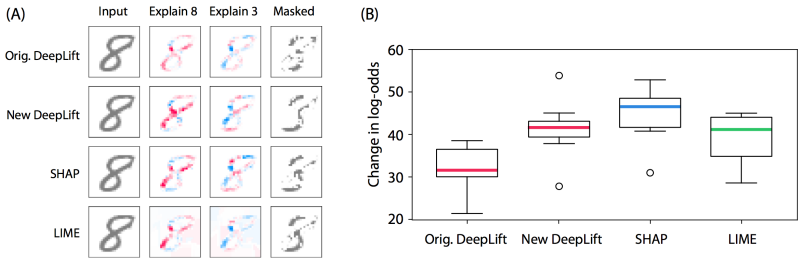
Abstract

"**Background**: Deep learning models are typically trained using stochastic gradient descent or one of its variants. These methods update the weights using their gradient, estimated from a small fraction of the training data. It has been observed that when using large batch sizes there is a persistent degradation in generalization performance - known as the"generalization gap" phenomena. Identifying the origin of this gap and closing it had remained an open problem.

Contributions: We examine the initial high learning rate training phase. We find that the weight distance from its initialization grows logarithmically with the number of weight updates. We therefore propose a "random walk on random landscape" statistical model which is known to exhibit similar "ultra-slow" diffusion behavior. Following this hypothesis we conducted experiments to show empirically that the "generalization gap" stems from the relatively small number of updates rather than the batch size, and can be completely eliminated by adapting the training regime used. We further investigate different techniques to train models in the large-batch regime and present a novel algorithm named "Ghost Batch Normalization" which enables significant decrease in the generalization gap without increasing the number of updates. To validate our findings we conduct several additional experiments on MNIST, CIFAR-10, CIFAR-100 and ImageNet. Finally, we reassess common practices and beliefs concerning training of deep models and suggest they may not be optimal to achieve good generalization."

Why we think it's interesting: The contributions of this paper fit into a recent stream of research, which explores generalization in deep learning models. This paper focuses on the performance gap between models that were trained using different batch sizes. The work is interesting because it has a direct practical impact. In contrast to contemporary best practices, larger batch sizes can be used for training without causing a significant decrease in generalization performance. An implication of this finding is the possibility to more easily achieve a higher degree of model parallelism in distributed and parallel optimization settings.

- **A unified approach in interpreting model predictions**

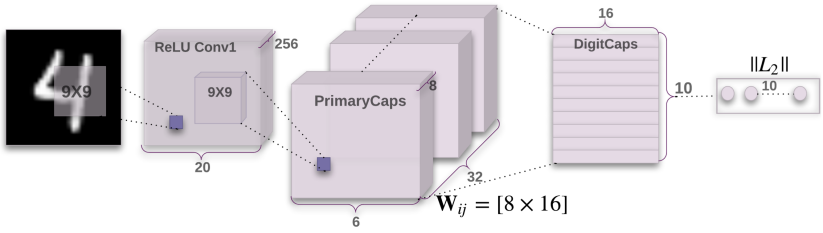


Abstract

"Understanding why a model makes a certain prediction can be as crucial as the prediction's accuracy in many applications. However, the highest accuracy for large modern datasets is often achieved by complex models that even experts struggle to interpret, such as ensemble or deep learning models, creating a tension between accuracy and interpretability. In response, various methods have recently been proposed to help users interpret the predictions of complex models, but it is often unclear how these methods are related and when one method is preferable over another. To address this problem, Lundberg et. al. present a unified framework for interpreting predictions, SHAP (SHapley Additive exPlanations). SHAP assigns each feature an importance value for a particular prediction. Its novel components include: (1) the identification of a new class of additive feature importance measures, and (2) theoretical results showing there is a unique solution in this class with a set of desirable properties. The new class unifies six existing methods, notable because several recent methods in the class lack the proposed desirable properties. Based on insights from this unification, we present new methods that show improved computational performance and/or better consistency with human intuition than previous approaches."

Why we think it's interesting: Explaining predictions made by deep learning models is highly relevant for the acceptance of those methods in various domains. This work is the first one defining a unified view on methods trying to attribute model decisions to certain input features. They use this underlying perception to further define methods, which belong to this family. However, the paper takes only subset of methods for model interpretations into account. Some of the more recent methods are not covered. Methods proposed in this paper have been integrated into two popular gradient boosting libraries (XGBoost and LightGBM).

- **Dynamic Routing of capsules**



Abstract

A capsule is a group of neurons whose activity vector represents the instantiation parameters of a specific type of entity such as an object or an object part. We use the length of the activity vector to represent the probability that the entity exists and its orientation to represent the instantiation parameters. Active capsules at one level make predictions, via transformation matrices, for the instantiation parameters of higher-level capsules. When multiple predictions agree, a higher level capsule becomes active. We show that a discriminatively trained, multi-layer capsule system achieves state-of-the-art performance on MNIST and is considerably better than a convolutional net at recognizing highly overlapping digits. To achieve these results the authors use an iterative routing-by-agreement mechanism: A lower-level capsule prefers to send its output to higher level capsules whose activity vectors have a big scalar product with the prediction coming from the lower-level capsule.

Why we think it's interesting: Capsules networks have drawn quite some attention within the machine learning community. They are especially interesting because they provide a new and meaningful way in which the interaction between layers of a neural network can be modeled. In many popular deep learning architectures, pooling layers combine and downsample inputs from a previous layer in a fixed way. Capsules networks instead provide a more dynamic and trainable form of pooling. Another key advantage is that for vision tasks, capsule nets are capable of capturing the spatial relationship between objects. However, training these networks in practice is still very inefficient compared to other architectures. More research is required in order to make them applicable in real-world scenarios.

- **Self-Normalizing Neural Networks**

Abstract

"Deep Learning has revolutionized vision via convolutional neural networks (CNNs) and natural language processing via recurrent neural networks (RNNs). However, success stories of Deep Learning with standard feed-forward neural networks (FNNs) are rare. FNNs that perform well are typically shallow and, therefore cannot exploit many levels of abstract representations. We introduce self-normalizing neural networks (SNNs) to enable high-level abstract representations. While batch normalization requires explicit normalization, neuron activations of SNNs automatically converge towards zero mean and unit variance. The activation function of SNNs are"scaled exponential linear units" (SELUs), which induce self-normalizing properties. Using the Banach fixed-point theorem, we prove that activations close to zero mean and unit variance that are propagated through many network layers will converge towards zero mean and unit variance – even under the presence of noise and perturbations. This convergence property of SNNs allows to (1) train deep networks with many layers, (2) employ strong regularization, and (3) to make learning highly robust. Furthermore, for activations not close to unit variance, we prove an upper and lower bound on the variance, thus, vanishing and exploding gradients are impossible. We compared SNNs on (a) 121 tasks from the UCI machine learning repository, on (b) drug discovery benchmarks, and on (c) astronomy tasks with standard FNNs and other machine learning methods such as random forests and support vector machines. SNNs significantly outperformed all competing FNN methods at 121 UCI tasks, outperformed all competing methods at the Tox21 dataset, and set a new record at an astronomy data set. The winning SNN architectures are often very deep. Implementations are available at: github.com/bioinf-jku/SNNs."

Why we think it's interesting: Batch normalization stands on shaky theoretical footing (as humorously pointed out by Ali Rahimi in his above-mentioned acceptance speech). Intuitively though, insuring the input to each network layer has zero mean and unit variance should improve the flow of gradients backwards through the network. In order to achieve this, the authors introduce a method that is less *ad hoc* than batch normalization. They place an activation function ("scaled exponential linear units" or SELUs) on the neurons, which tend to converge towards normalized outputs. This allows effective training of deep networks with nice results.

Conclusion

NIPS remains the top conference for machine learning by keeping very high quality standards in terms of papers, talks and workshops, as well as covering a very broad range of domains. The conference has seen a record number of attendees and sponsors in Long Beach, CA , as machine learning is gaining traction in a variety of domains. We expect this trend to continue in following years while machine learning integrates into our daily lives.