YES Workshop 2019

Understanding Deep Learning: Generalization, Approximation and Optimization

Tutorial Speakers:

Johannes Schmidt-Hieber

Lecture 1) Statistical theory for shallow networks and advantages of additional layers

We start with the universal approximation theorem and discuss several proof strategies that provide some insights into functions that can be easily approximated by shallow networks. Based on this, a survey on approximation rates for shallow networks is given. It is shown how this leads to estimation rates. In the lecture, we also discuss methods that fit shallow networks to data.

Why are deep networks better than shallow networks? We provide a survey of the existing ideas in the literature. In particular, we study localisation properties of deep networks and discuss the Kolmogorov-Arnold representation theorem.

Lecture 2) Statistical theory for deep ReLU networks

We outline the theory underlying the recent bounds on the estimation risk of deep ReLU networks. In the lecture, we discuss specific properties of the ReLU activation function that relate to skip connections and efficient approximation of polynomials. Based on this, we show how risk bounds can be obtained for sparsely connected networks.

Lecture 3) Comparison to other statistical methods and open problems

There is some literature comparing deep vs. shallow networks. The statistical approach allows us also to provide a theoretical comparison of the performance of deep networks to other well-established methods such as wavelet methods and MARS (multivariate adaptive regression splines). The second part of the third lecture is devoted to outline future challenges in the field. We describe important steps needed for the future development of the statistical theory.

Peter Bartlett

Generalization in deep networks. I, II: we study how the performance of prediction rules on training data compares to their predictive accuracy for classification and regression problems. We examine how this depends on notions of complexity of the prediction rules, such as Vapnik-Chervonenkis dimension and Rademacher averages, and how these quantities can be bounded for neural networks. Deep networks raise some novel challenges, since they have been observed to perform well even with a perfect fit to the training data. In an effort to understand the performance of interpolating prediction rules, we present

some preliminary results for the simple case of linear regression, and highlight the questions raised for deep learning.

Optimization in Deep Residual Networks: we consider properties of deep residual networks that are relevant to optimization. For networks that compute near-identity maps at each layer, we find that their representational power improves with depth, and that the functional optimization landscape has the desirable property that stationary points are optimal. We consider implications for optimization in deep linear networks, showing how the success of a family of gradient descent algorithms that regularize towards the identity function depends on a positivity condition of the regression function.

Nati Srebro – Learning with Underdetermined Non-Convex Models (aka Deep Learning)

In this three-part tutorial we will begin by reviewing the foundational principles of machine learning, and then see how they apply to learning multi-layer models, aka “deep learning”. We will then try to understand the appeal, and also the mysteries, of deep learning. In particular, we will highlight the crucial role of implicit or explicit regularization in parameter space, and what this induces in function space.

Invited Speakers

Max Welling - From Graphical Models to Deep Learning: a Synthesis.

Graphical models have been the workhorse of machine learning for over a decade. Until deep learning took over as the new paradigm. But graphical models (GM) and deep learning (DL) should be considered complementary technologies. Graphical models are a great tool to express prior knowledge, including causal relationships. They provide a much more explainable and statistically well founded paradigm for building models and reasoning about random variables. Deep learning on the other hand provides a powerful tool for "black box" prediction on raw inputs. A natural question is how to fruitfully combine these paradigms. In this talk I will discuss the various efforts we have pursued in AMLAB to achieve this GM-DL synthesis. We discuss variational auto-encoders, recurrent inference machines and graphical recurrent inference networks as examples of an emerging unified framework.

Julien Mairal - Group Invariance and Stability to Deformations of Deep Convolutional Representations

In this work, we study invariant properties of convolutional neural networks, their stability to image deformations, and their model complexity from a kernel point of view. This is achieved by designing a multilayer kernel representations for deep neural networks and by studying the geometry of the corresponding reproducing kernel Hilbert space. We show that the signal representation is stable and that models from this functional space, such as a large class of convolutional neural networks with homogeneous activation functions, may enjoy the same stability. In particular, we study the norm of such models, which acts as a measure of complexity that controls both stability and generalization. This is a joint work with Alberto Bietti.

Taco Cohen - Gauge Fields in Deep Learning

Gauge field theory is the foundation of modern physics, including general relativity and the standard model of physics. It describes how a theory of physics should transform under symmetry transformations. For instance, in electrodynamics, electric forces may transform into magnetic forces if we transform a static observer to one that moves at constant speed. Similarly, in general relativity acceleration and gravity are equated to each other under symmetry transformations. Gauge fields also play a crucial role in modern quantum field theory and the standard model of physics, where they describe the forces between particles that transform into each other under (abstract) symmetry transformations.

In this work we describe how the mathematics of gauge groups becomes inevitable when you are interested in deep learning on manifolds. Defining a convolution on a manifold involves transporting geometric objects such as feature vectors and kernels across the manifold, which due to curvature become path dependent. As such it becomes impossible to represent these objects in a global reference frame and one is forced to consider local frames. These reference frames are arbitrary and changing between them is called a (local) gauge transformation. Since we do not want our computations to depend on the specific choice of frames we are forced to consider equivariance of our convolutions under gauge transformations. These considerations result in the first fully general theory of deep learning on manifolds, with gauge equivariant convolutions as the necessary key ingredient.

Sander Bohté – The biology of deep learning

As some may recall, deep learning and neural networks are bio-inspired, in the sense that some of their functioning is inspired by beliefs about how the neurons in the brain conspire to transform perception into action. What do we at present know about deep learning in the brain? What can we learn from it? Is it relevant? These are some of the topics I will touch upon, discussing networks of spiking neurons, models of biologically plausible deep learning, and the relation to the emerging field of neuromorphics.

Contributed Speakers

Mariia Vladimirova - Understanding priors in Bayesian neural networks at the Unit level

We investigate deep Bayesian neural networks with Gaussian priors on the weights and ReLU-like nonlinearities, shedding light on novel sparsity-inducing mechanisms at the level of the units of the network, both pre- and post-nonlinearities. The main thrust of the work is to establish that the units prior distribution becomes increasingly heavy-tailed with depth. We show that first layer units are Gaussian, second layer units are sub-Exponential, and we introduce sub-Weibull distributions to characterize the deeper layers units. Bayesian neural networks with Gaussian priors are well known to induce the weight decay penalty on the weights. In contrast, our result indicates a more elaborate regularization scheme at the level of the units. This result provides new theoretical insight on deep Bayesian neural networks, underpinning their natural shrinkage properties and practical potential.

Anastasia Botovykh - Understanding generalisation in noisy time series forecasting

In this presentation we study the loss surface of neural networks for noisy time series forecasting. We aim to gain insight into the effects of deep versus wide networks on the loss surface and the structure of the critical points. In extrapolation problems for noisy time series, neural networks, due to their overparametrization, tend to overfit and the behavior of the model on the training data does not measure accurately the behaviour on unseen data due to e.g. changing underlying factors in the time series. Avoiding overfitting and finding a pattern in the data that persists for a longer period of time can thus be very challenging. In this talk we quantify what the neural network has learned using the structure of the loss surface of multi-layer neural networks. We study the role that the optimization method plays in the generalisation capabilities and gain insight into which minima are able to generalise well based on the spectrum of the Hessian matrix and the smoothness of the learned function with respect to the input. Based on the structure of the time series we quantify how to make the trade-off between the complexity of the learned function and the ability of the function to fit the data well so that the network can generalise well.

Benjamin Bloem-Reddy - Probabilistic symmetry and invariant neural networks

In an effort to improve the performance of deep neural networks in data-scarce, non-i.i.d., or unsupervised settings, much recent research has been devoted to encoding invariance under symmetry transformations into neural network architectures. We treat the neural network input and output as random variables, and consider group invariance from the perspective of probabilistic symmetry. Drawing on tools from probability and statistics, we establish a link between functional and probabilistic symmetry, and obtain functional representations of probability distributions that are invariant or equivariant under the action of a compact group. Those representations characterize the structure of neural networks that can be used to represent such distributions and yield a general program for constructing invariant stochastic or deterministic neural networks. We develop the details of the general program for exchangeable sequences and arrays, recovering a number of recent examples as special cases.
This is work in collaboration with Yee Whye Teh. https://arxiv.org/abs/1901.06082

Wouter Koolen - Active Learning in the age of Deep Learning

Deep neural networks are successfully applied in passive learning settings, where a labelled data set is readily available. In this talk we look at active learning problems, where the learning system controls data acquisition, as well as reinforcement learning, where the actions of the learning algorithm influence the state of the world. We fist introduce the problem setting, and argue that active learning holds the promise of greatly reducing sample complexity. We then zoom into one successful example, AlphaZero, review its architecture and training procedure, and highlight the roles of exploration, planning and curve fitting. We complement the picture with recent results in pure exploration for stochastic bandits, where asymptotically optimal algorithms are starting to become available for a variety of problems, and look at the lessons for future integrated systems.

Lukas Szpruch - Deep Learning through the lens of Mean-Filed Control Problem

In this work, we will demonstrate that one can view training of deep neural network as sampling from a probability distribution over the parameter space. In the case of one hidden layer, it has been observed in the literature that the original finite-dimensional non-

convex optimisation problem becomes convex at the cost of working on infinite dimensional space of probability measures. Nonetheless using techniques from the theory of Optimal Transport and gradient flow on the Wasserstein space in particular, one can provide theoretical guarantees for global converges for the general optimisation algorithm. Gradient flow evolution equation, also known as McKean-Vlasov eqautions, can be then projected on finite dimensional space using a system of interacting particles in the spirit of propagation of chaos. This top-down approach is very fruitful as it allows to recover stochastic gradient algorithms used in practice but also paves the way to new algorithms for training neural nets. I shed some light on how the above methodology can be used to train multilayer networks.