Mathematical
Institute

Summary of theories of deep learning material covered

Theories of Deep Learning: C6.5,
Lecture / Video 15

Prof. Jared Tanner
Mathematical Institute
University of Oxford

Oxford
Mathematics

- Structure of a deep net as repeated affine transforms and non-linear activations.
- Introduction to LeNet-5 with convolutional and fully connected layers.
- MNIST as an example of small dataset, along with the more complex imagenet dataset.
- Discussion of availability of computational resources and optimisation algorithms.

- Telgarsky 15' sawtooth function giving example function with exponentially many maxima as a function of depth, but linear in width.
- Yarotsky 16' extension of the sawtooth to generate local polynomial approximations within ϵ needing $\log(1/\epsilon)$ depth.
- Poggio et al. 17' tree structure for approximation rate at the low effective dimensionality.
- Hein et al. 05' showing that MNIST, for example, has low effective dimension.

UNIVERSITY OF OXFORD
Mathematical Institute

- Glorot et al. 10' observation of pre-activation hidden layers being approximately Gaussian and normalizing to have constant variance with depth.
- Poole et al. 16' quantifying the convergence of the Gaussian variance through a computable recursion relation and developing a theory for the correlation between inputs dependence on $(\sigma_w, \sigma_b, \phi(\cdot))$. Controlling geometric collapse or instability through selecting $(\sigma_w, \sigma_b, \phi(\cdot))$ according to derived formulae.
- Pennington et al. 18' introduced the edge of chaos and connecting the Poole et al. work with exploding and vanishing gradients. Random matrix theory to derive moments of the spectra of the DNN.

UNIVERSITY OF OXFORD
Mathematical Institute

- Foundational theory on stochastic gradient descent (SGD), making use of back-propagation and mini-batches to improve scalability of the algorithms.
- Reducing the gap between the value of a global minimizer and SGD through decreasing stepsize, increasing batchsize, or other variance reduction methods.
- Accelerating through momentum and diagonal scaling.
- Ward et al. AdaGrad scalar diagonal scaling to have reliable training over a wide range of stepsize initializations.

University of Oxford
Mathematical Institute

- Venturi et al. 16' showed the minimizers of the loss landscape can have paths between them where the loss landscape remains of a similar value, nearly simply connected.
- Pennington et al. 17' showed how the distribution of eigenvalues of the loss landscape's Hessian can be computed as a function of the number of trainable parameters compared to the amount of data available. With enough data, and when close to a minimizer, then the landscape is locally convex.
- Li et al. 18' illustrated how the loss landscape width near minimizers is impacted by training batch-size, and how ResNets can improve the landscape characteristics.
- Loffe 15' introduced batchnorm to have trainable bulk scaling to aid optimisation.

- Filters for early layers of CNNs show characteristic high-dimensional wavelet like structure.
- Deeper layers combine such filters and give greatest response for more structured inputs, leading to "memory" where some units in the CNN are maximized by objects within training classes.
- Such structure helps explain the efficacy of transfer learning.
- Mallat 12' introduced the Scattering Transform, which is a deep transform that only learns the final layers; activations are hand crafted to encourage desired invariants such as translation.

- Goodfellow et al. 14' introduced the generative adversarial network (GAN) structure using reversed DNNs to generate data characteristic of a training dataset.
- Bau et al. 20' Filters analogous to those in CNNs on natural images are learned for generating natural images. They can be modified in order to influence expected properties of the generated data, such as colour or frequency of objects such as trees or doors in buildings.
- Moosavi-Dezfooli et al. 16' introduced DeepFool to effectively generate adversarial misclassification.

- Engstrom et al. 18' showed that natural actions on objects, such as rotation or translation can also be used to generate misclassification; showing inherent lack of robustness in typical DNNs.
- Wong et al. 17' demonstrated that one can adapt the training to have certificates which ensure that the network is provably robust in some circumstances.
- Gopalakrishnan et al 18' showed how sparsifying a DNN can improve robustness, explainable by reducing an upper bound on the DNN's Lipshitz constant.
- Autoencoders and VAEs as alternative architectures that can be built from DNNs for tasks such as denoising or explainable generative data.

- Exponential approximation ability with depth is well understood.
- Algorithms exist to effectively train DNNs and their variants, overcoming disadvantages that accompany the exponential nature of depth.
- DNNs are observed to generalize well, but theory is lacking to show a benefit of depth in terms of generalization; such a result may not exist, or may require a DNN that is regularized to show that with a fixed amount of trainable parameters, that depth has generalization benefits over width.
- This leads us to consider the amount of trainable parameters needed; width and depth are needed, but generate undesirably large amounts of trainable parameters.