DOI:10.1145/3446773

To view the accompanying paper, visit doi.acm.org/10.1145/3446776 rh

Technical Perspective
Why Don't Today's Deep Nets Overfit to Their Training Data?

By Sanjeev Arora

THE FOLLOWING ARTICLE by Zhang et al. is well-known for having highlighted that widespread success of deep learning in artificial intelligence brings with it a fundamental new theoretical challenge, specifically: *Why don't today's deep nets overfit to training data?* This question has come to animate the theory of deep learning.

Let's understand this question in context of *supervised learning*, where the machine's goal is to learn to provide labels to inputs (for example, learn to label cat pictures with "1" and dog pictures with "0"). Deep learning solves this task by training a net on a suitably large training set of images that have been labeled correctly by humans. The parameters of the net are randomly initialized and thereafter adjusted in many stages via the simplest algorithm imaginable: gradient descent on the current difference between desired output and actual output.

At the end of training, one usually finds that labels assigned by the net on the training images are mostly or entirely correct. Does this mean the net can be used to correctly label other pictures we will find on the Internet? Not necessarily. It is conceivable the net learned to correctly label just the training pictures, and no others. In other words, it could have *overfitted* to the training data. It is customary to check for this using a *holdout* set of training data that was left unused during training. The assumption underlying this methodology is that training data consists of independent samples from a fixed distribution, and we desire a net that gives correct labels to most images of the entire distribution. A simple probability concentration bound shows that performance on the holdout set is predictive—up to some well-defined error bars—of performance on the unseen images from the same distribution.

Received wisdom has it that overfitting happens if the net is *too expressive*, that is, has sufficient number of layers, and parameters per layer, than it is capable of expressing arbitrarily complicated mappings from inputs to 0/1 labels. To avoid overfitting, one should use a model that cannot "achieve more complicated functions than necessary." This philosophical principle is called *Occam's Razor* and related to the reasons why we prefer simpler scientific theories to complicated ones.

Decades of work in theory of machine learning and statistics has yielded measures of model complexity ranging from the old VC dimension and Rademacher complexity to more modern norm-based measures. This theory suggests that during training one must add a *regularizer* term to the training objective that penalizes models with a high measure of complexity.

Modern deep nets have turned out to confound this intuitive framework of regularizers. As the paper shows, it is possible to train nets with 50 million parameters using no regularizers on only 10,000 training examples. Surprisingly, no significant overfitting happens.

The extensive experiments detailed in the paper serve to deepen the mystery of this lack of overfitting. The experiments involve training nets on randomized/nonsensical versions of standard images datasets—the most benign being randomization of labels and more extreme being using random collections of pixels as images and random labels. Current deep nets—even with standard training and regularizer—are capable of achieving a good fit on these nonsensical datasets, which shows that these nets are capable of expressing very complicated functions. In particular, the experiment of fitting a net on images with random labels shows that a traditional measure, Rademacher complexity, is high for the deep net architecture.

Subsequent work has explored the authors' suggestion that the training algorithm (a variant of gradient descent) plays a powerful role in how overfitting is avoided. Many new measures have been defined to measure the "effective number of parameters" of a trained net. Several of these measures were reported to correlate with good generalization. However, a recent extensive study[2] suggests this correlation is pretty weak and we still don't have a conclusive idea of why overfitting does not happen.

Another intriguing direction that has led to a flurry of papers is theoretical understanding of extreme over-parametrization. Since over-parametrization does not seem to hurt deep nets, it is natural to wonder if one can take it to the extreme. Recent work has analyzed the infinite limit: take a finite net and allow its width (= number of nodes for fully connected layers, and number of channels for convolutional layers) to go to infinity. This is the wonderful world of *Neural Tangent Kernels* or NTK.[1] Perhaps some of these new ideas will appear in the pages of *Communications* in future. Kudos to Zhang et al. for writing a paper that led to all this interesting follow-up work! C

References

1. Jacot, A. et al. Neural tangent kernel: Convergence and generalization in neural networks. In *Proceedings of the 32nd NeurIPS Conf.* (Montreal, Canada, 2018).
2. Neyshabur, B. et al. Towards understanding the role of over-parametrization in generalization of neural networks. In *Proceedings of ICLR*, 2019.

Sanjeev Arora is the Charles C. Fitzmorris Professor of Computer Science at Princeton University, Princeton, NJ, USA.