Capacity, Regularization, and Generalization in Deep Neural Networks

Daniel Ehrlich

April 6, 2018

1 Reading Materials

- Zhang, Chiyuan et al., "Understanding Deep Learning Requires Rethinking Generalization" arXiv:1611.03530 (2016)
- Arpit, Devansh et al., "A Closer Look at Memorization in Deep Networks" arXiv:1706.05394 (2017)
- Morcos, Ari et al., "On the Importance of Single Directions for Generalization" arXiv:1803.06959v1 (2018)

2 Abstract

Generalization is the ability of a model trained on a subset of examples to correctly perform on new examples not seen during training. Statisticians have long observed a strict trade-off between training and test (generalization) errors. While initially the two forms of errors tend to go down as a function of model complexity, eventually the models will tend to "overfit" the noise in the training data and the test error will begin to rise.

Traditional theories of generalization in statistical learning have combated overfitting with two elements, limitations in the expressivity of a given model family and external regularization. Models with limited capacity are only able to focus on the most robust sources of signal and therefore tends to identify solutions less dependent on noise. Regularization on the other hand explicitly penalizes model complexity and overfitting, commonly with an L1 or L2 penalty in the parameter space. Several recent papers, however, have questioned whether capacity and regularization have the same impact on deep learning.

Zhang et al., demonstrate that despite the strong generalization performance of networks like AlexNet and InceptionNet, parameter size and model complexity are too large to act as a constraint on overfitting. The authors generated random labels for the CIFAR-10 dataset, and trained each network to do the random classification problem. Despite no systematic signal, the networks had sufficient capacity to simply memorize each image-label pair. Further they demonstrate that adding explicit regularization (drop-out, weight decay and batch-norm) had no meaningful impact on final training accuracy on the random label set. From this they conclude that in practical deep learning settings, capacity and regularizations do not constrain overfitting.

In a follow up paper, Arpit et al. expand but also question this initial finding. Counterintuitively they demonstrated that increasing network size and capacity made deep networks more generalizable, in stark contrast to the traditional theory that reducing capacity would improve generalization. However, they also demonstrate systematic differences between training dynamics for random and structured data. Importantly, real data showed bias in which "patterns" and examples were classified early in training, while random data did not. In addition they demonstrated that the mean loss over time for real data tended to be a function of a small selection of "indicative" examples, while in random data all examples had a large impact on loss. Taken together this implies that in the real data case it is unlikely that memorization of individual examples is playing a large role, even if the network is capable of utilizing memorization in the random case.

Finally, in the most recent paper Morcos et al. seek to understand why some networks generalize well while others may not. They use an experimental perturbation method to test the reliance of different instantiations of a trained neural network on single dimensions in activity state space. They show that networks that are less reliant on single directions tend to generalize better to test data, potentially explaining the impact of drop-out and batch norm on regularization in deep networks. Further this implies that the distributed representations implicit in deep learning architectures may be acting as an independent regularizer.

3 Spotlight Question

What factors other than network design and explicit regularization may play a role in the generalization performance of deep neural networks?