Your 598 ID: 5107

Title of paper: "Stronger generalization bounds for deep nets via a compression approach" by Arora, Ge, Neyshabur, Zhang.
What is their primary result? Deep neural networks generalize very well despite having far more parameters than the samples they are trained on. This paper introduces a simple compression-based framework, which yield good generalization bounds (on par with empirical results).
Why is this important? Previous methods of analyses (via PAC-Bayes and Margin) do not yield bounds better than naive parameter counting.
What are their key ideas? The main idea introduced by the authors is the *compression framework* they introduce in Section 2 of the paper and follows from the following observation. Suppose f is a classifier with m parameters that incurs very low empirical loss. If we can compute a classifier g with discrete trainable parameters much less than m and which incurs a similar loss on training data as f, then g incurs low classification error on the full distribution.

The authors introduce an algorithm for computing the weights of such a compression for a fully connected network (subject to some mild constraints on noise stability introduced in Section 3) and show that such a compression achieves good generalization bounds.

Lastly, the authors sketch how to compress convolutional networks in Section 5.
How did they train and evaluate it? The authors compute the generalization bound for Theorem 5.1 on a VGG-19 architecture and an AlexNet in the multi-class classification task on the CIFAR-10 dataset. They optimize it with SGD with a minibatch size of 128, weight decay of $5 \cdot 10^{-4}$, momentum 0.9, and initial learning rate of 0.05 but decayed by a factor of 2 every 20 epochs. Moreover, they use dropout on fully connected layers. The networks are trained for 299 epochs and the final VGG-19 network achieves 100% on training and 92.45% on validation.

The authors identify four properties (layer cushion, interlayer cushion, contraction, and interlayer smoothness) introduced in Section 3 which contribute to noise stability and demonstrate empirically that their trained networks satisfy these.
Did they implement something? The authors implement an algorithm for computing compressed classifier.
Grader's 598 ID: