Summary of Deep Double Descent

Candace McKeag

October 24, 2020

Abstract

In this brief summary of the paper "Deep Double Descent: Where Bigger Models and More Data Hurt" by Nakkiran et al. (2019), we discuss the paper's key points, novelty, terms, experiment results, and relation to neural networks and STATS 231A.

Summary

In the paper "Deep Double Descent: Where Bigger Models and More Data Hurt" by Nakkiran et al. (2019), the authors challenge two widely-conventional wisdoms of the classical statistics realm: that larger models are worse, and that more data is always better. They introduce a generalized double descent hypothesis consisting of three types of regimes during model training: under-parameterized, over-parameterized, and critically parameterized.

Double descent is a phenomenon that occurs when as model size increases, the performance first gets worse and then gets better. When initially increasing model complexity, the test error exhibits a U-like curve similar to that described by the bias-variance trade-off, a common topic in classical statistics. This is the under-parameterized region. At the right side of the U curve, the test error peaks in the critical region or interpolation threshold. As model complexity is further increased, the paper describes situations where the test error actually continues to decrease, thus challenging the belief that larger models are worse. There are two types of double descent phenomena discussed/proposed in the paper: epoch-wise and model-wise.

To demonstrate epoch-wise double descent, the authors keep the model architecture fixed and increase training time. They found that the U-like curve related to the bias-variance trade-off appeared in the underfitting stage, but improved once the complexity measure has surpassed the sample size. They saw a peak in test performance once the models reached close to 0 training error. In sufficiently large models, they found that at first the test error decreased, then increased near the interpolation threshold, and then decreased again. This is the double descent behavior, which was found to be robust across optimizer variations and learning rate schedules.

Contrary to demonstrating epoch-wise double descent, experimenting with model-wise double descent involves fixing the number of optimization steps, training to completion, then studying the test error as the model size increases. A peak in test error was found to systematically occur at the interpolation threshold. The intuition behind this is that for

model sizes that lie at the interpolation threshold, there is effectively only one model that fits the training data. Because there is only one model that barely fits, it is very sensitive to noise in the training set, which can result in high test error. For over-parameterized models, there are many interpolating models that can fit the training set, and thus they see lower test error.

The authors also introduce a complexity measure called the effective model complexity (EMC), which is defined as the maximum number of samples on which the model can achieve close to zero training error. This measure depends on the data distribution and true labels, classifier architecture, training time, and training procedure. The point where the EMC matches the number of samples is the transition point between under- and over-parameterization. This point is where the test error peaks. In the experiments, the authors found that this point can be shifted to the right when the number of samples increases, implying that adding more data is worse in these situations.

The authors discuss a concept called sample non-monotonicity, seen in the experiments. They fixed the model architecture and training procedure while varying the number of training samples. In their experiments, they observed distinct test behavior in the critical regime: a long plateau region in which adding more data might hurt. By increasing n, they saw that the training procedure might switch from being effectively over-parameterized to effectively under-parameterized. Increasing n could also cause the area under the error curve to shrink, and/or shift the curve to the right, thus increasing the model complexity at which the test error peaks. These effects could either combine or cancel each other out, both of which would not result in better test performance.

The main ideas behind this paper are to demonstrate that 1) double descent can lead to a regime where training on more data leads to worse test performance, and 2) there are model settings in which over-parameterization and larger models yield lower test error. The novelty of the paper lies in its extension of the idea of double descent beyond the number of parameters, and rather to include the training procedure under a unified notion of EMC. The experiments performed and demonstrated also provide a rigorous exhibition of double descent for modern modeling tasks. I think that this paper offers a very interesting counter-argument to some widely-held beliefs in the fields of machine learning and statistics. Specifically, this paper supports the perhaps counter-intuitive observation that neural networks are extremely powerful and accurate despite often having millions of parameters. Relative to STATS231A, I think this paper can help us understand the importance of finding balance within the multitude of parameters required in building complex neural networks, and when it is appropriate to over-parameterize a model.

References

[1] P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, and I. Sutskever, *Deep Double Descent: Where Bigger Models and More Data Hurt*, International Conference on Learning Representations, 2020.