ECE472 – Quiz 8

Jonathan Lam

November 11, 2020

1. *Make an argument about super-convergence as it relates to epoch-wise double-descent.* (arXiv:1708.07120, arXiv:1912.02292)

   Double-descent is a phenomenon related to the "effective model complexity" (EMC) of a deep neural network, which can roughly be thought of as the number of samples the model can correctly classify (a measure of "capacity"), as well as the number of training samples. The authors show that EMC is a function both of model complexity (more parameters means a greater network capacity) and number of training epochs (training more epochs intuitively means being able to classify more samples correctly). The phenomenon is that for a low EMC, there is a "U"-shaped curve as the model is under-parameterized, then begins to fit, and then overfits; after a certain point the model begins a "second descent" as the model generalizes again with higher EMC. The increase in test error when overfitting follows the conventional wisdom that the model becomes overly-specialized to the input data, but this logic seems to fail with the second descent. This overfitting occurs when the EMC is close to the number of training samples.

   (A possible intuition for this is that as the EMC gets close to the number of input samples, then it almost exactly models the data, and tries to fit it very closely; the model is a perfect fit for the data. As the EMC increases, then the model becomes "over-parameterized"; i.e., many different models can fit the training data, and so the model can generalize better. (E.g., given 1001 points (none of which lie on the same vertical line), you can fit a 1000th-degree polynomial perfectly in exactly one way; with a higher-degree polynomial you can still fit these points but in infinitely may ways, and possibly with a smoother interpolation between the test points.))

   Superconvergence is related to a training method that uses cyclical (or "1cycle") learning rates. This is in contrast to the conventional global or monotonically-decreasing piecewise-constant learning rate. This provides a regularizing effect, and may be well-suited to the specific type of overfitting that occurs with double-descent. Cyclical learning rates (and the "1cycle" method, in which the learning rate is kept high until near the end of training) discourage the premature convergence that happens when the EMC approaches the number of training samples. Other forms of regularization also slow convergence, but they slow the increasing of the EMC as well; superconvergence makes the EMC increase even faster with its large learning rates, potentially increasing the EMC quickly past the double-descent hump. Similarly to regularization, learning rate optimizers like Adam may try to prematurely converge when the EMC approaches the number of training samples by slowing learning, and this will also make it harder to overcome the double-descent hump.