Measuring the simplicity of neural networks as a function of overparametrization

Elizaveta Demyanenko, Luca Saglietti, Enrico Malatesta
Bocconi University, Milan, Italy

Understanding generalization properties is a crucial part of neural networks analysis. At odds with the standard bias-variance trade-off described by statistical learning theory, recent works demonstrated the existence of a double descent phenomenon for the generalization error of neural networks, highlighting a variety of settings where the test performance of these models can improve above the interpolation threshold. In the present work, we aim at building a link between the double descent phenomenon and the sensitivity of the function represented by neural networks. In particular, we characterize the susceptibility of the network with respect to the input features and their degree of ”influence” on the output of the model. To exactly quantify this influence we employ the mean dimension (MD), a metric developed in the context of boolean function analysis. The MD of a pseudo-boolean function (taking an n-dimensional binary input and producing a scalar output) indicates the average dependency of the function on multipoint input correlations, and can be obtained from the Fourier coefficients in the decomposition of the function in the basis of 2^n polynomials, being products within all the possible sets of the features. Operationally the MD can be estimated by measuring a weighted sum of feature influences divided by the total variance of the function. In this way, the MD of a neural network can be used to measure the “simplicity” of the represented function from its response to variations in the input. We find that, as the degree of overparametrisation of the network is increased, the MD reaches an evident peak at the interpolation point, in perfect correspondence with the double descent of the generalization error, and then slowly approaches a low asymptotic value. We analyze this phenomenon for different model classes and training setups. Moreover, we demonstrate that models more robust to adversarial attacks exhibit lower mean dimension, and on the contrary, adversarially initialised models tend to show a higher mean dimension in our experiments.