EECS 6322 Week 12, Paper 2
Deep Double Descent: Where Bigger Models And More Data Hurt

Shivani Sheth (Student #: 218011783)
shivs29@yorku.ca

I. Summary and contributions

The paper brings forward a phenomenon, known as double descent, that is observed in deep learning models. The authors observe through a series of experiments that the performance of deep learning models first gets worse as the model size increases, but then gets better. The paper proposes a new complexity measure, known as the effective model complexity to discuss the generalized double descent in deep learning models and they show that the phenomenon depends on various factors other than the model size, such as the number of training epochs and the noise in the training dataset. Finally, the authors also show that an increase in the number of training samples actually hurt the performance of the model, as contrary to the popular belief.

The fundamental concept of bias-variance trade-off states that the higher complexity models have a lower bias but a higher variance, due to which after a certain threshold, the models overfit the data leading to a decrease in the model performance and an increase in the test error. Hence, according to this conventional theory, larger models perform worse after a certain threshold. However, neural networks in practice contradict this theory. Neural networks have millions of parameters that can easily fit the random noise in the data and yet they perform better than the smaller models in many tasks. Hence, in practice, larger models are found to be better. With respect to data, both domains agree that more data should be better for the performance of the models. These conventional concepts are challenged by the paper and through empirical results are proved to be true only in the under-parameterized regime, where the model complexity is small as compared to the number of samples.

The authors propose an effective model complexity (EMC) measure and provide a general notion of double descent that is a function of the EMC measure. The EMC of a training procedure is defined as the maximum number of samples on which it can achieve close to zero training error, and depends on the data distribution, the architecture of the classifier, and the training procedure. The performance of the model is observed to follow a U-like curve in the underfitting stage, when the EMC is smaller than the number of samples, and then improves with training time once the EMC increases as compared to the number of samples. The performance of the model with respect to the number of training samples also shows that when there is a transition from the under to over parameterization and the EMC matches the number of samples, the test error of the model is at its peak. An increase in the number of training sample shifts this peak to the right and hence in some cases can result in a situation where more data is worse for the performance of the model.

The effective model complexity of a training procedure with respect to a distribution is defined as the maximum number of samples n on which the training procedure achieves an average of 0 training error. This is given by $EMC_{\mathcal{D},\epsilon}(\mathcal{T}) := max\{n|E_{S\sim\mathcal{D}^n}[Error_S(\mathcal{T}(S))] \leq \epsilon\}$, where parameter $\epsilon > 0$ and $Error_S(M)$ is the mean error of model M on training samples S. Based on the EMC, the authors propose a hypothesis that divides the performance of a model into 3 regimes. First is the under-parameterized regime, where the effective model complexity is smaller than the number of training samples. In this case, an increase in the effective model complexity decreases the test error. Second is the over-parameterized regime, where the effective model complexity is larger than the number of training samples. In this case, an increase in the effective model complexity decreases the test error. Finally, third is the critically parameterized regime, where the effective model complexity is approximately equal to the number of training samples. Here, an increase in the effective model complexity may increase or decrease the test error.

The authors show that the double descent curve and hence the model's performance also changes with respect to changes in the optimization algorithms, model size, number of epochs, number of training samples, and the label noise in the training samples. In model-wise double descent, the paper studies the test error of models of increasing size, for a fixed large number of optimization steps and shows realistic settings in which bigger models are worse. In epoch-wise double descent, the paper studies the test error of a fixed, large architecture over the course of training and shows that training longer can correct overfitting. In sample-wise non-monotonicity, the paper studies the test error of a fixed model and training procedure, for a varying number of training samples and shows that for a fixed architecture and training procedure, more data actually hurts. Finally, with respect to label noise in the training data, the paper observes that all forms of double descent most strongly occur in settings that have label noise in the train set, which is often the case when collecting train data in the *real-world.*

The study conducted by the authors include three families of model architectures which are ResNets, standard CNNs, and Transformers. The models use a cross-entropy loss with the Adam and SGD optimizers. The label noise of probability p is given by the number of samples which have the correct label with probability $(1-p)$, and a uniformly random incorrect label otherwise. Experiments on the model-wise double descent showed that the critical region exhibits distinctly different test behavior around the interpolation point and there is often a peak in test error that is more prominent with label noise. It also showed that all modifications which increase the interpolation threshold, such as label noise, data augmentation, and increasing the number of train samples, also correspondingly shift the peak in test error towards larger models. The intuition that the authors provide for the phenomena with respect to the model size is that at the interpolation threshold, only one model fits the train data that is very sensitive to the train set and/or model mis-specification. Hence, it also fits the noisy/ mis-specified labels, which destroy its global structure, and result in a high test error. On the contrary, for the over-parameterized models, many interpolating models fit the train set and SGD finds the one that memorizes/ absorbs the noise while still performing well on the distribution.

The experiments on epoch-wise double descent show that increasing the train time increases the effective model complexity, and thus a sufficiently large model transitions from under to over parameterized through the course of training. Similarly, experiments on the sample wise non-monotonicity explore the critical regime by varying the number of train samples n, and show that by increasing n, the same training procedure $\mathcal{T}$ can switch from being effectively over-parameterized to effectively under-parameterized. It also shows that for a given regime of model sizes, more data actually hurts the test performance of the model.

II. Strengths

The paper proposes a generalized concept of double descent that shows that models and training procedures exhibit atypical behavior when their Effective Model Complexity is approximately equal to the number of training samples. It shows that the phenomenon is robust to the dataset, architecture, training procedures, and demonstrates a model-wise double descent to characterize the regime where bigger models can perform worse. The authors also show that the double descent phenomenon can lead to a regime where training on more data leads to worse test performance.

III. Weaknesses

The hypothesis proposed by the authors does not have a principled way to choose the parameter ϵ and does not provide a formal specification for the "sufficiently smaller" and "sufficiently larger" model complexities. It also does not provide sufficient theoretical grounding on the effect of the training procedure and distribution on the width of the critical interval.

IV. Correctness

The claims and empirical methodology of the paper are correct and consistent with the phenomenon of double descent on multiple experiments that the authors demonstrate. The hypothesis proposed by the authors can be seen on a variety of architectures such as ResNets, standard CNNs, Transformers, on a variety of datasets such as CIFAR10, CIFAR100, IWSLT'14, and two different training algorithms which are SGD and Adam. Hence, the phenomenon is consistent with the claims of the authors which can also be seen by deep learning practitioners.

V. Clarity

The paper is well written with sufficient graphs and figures to represent the results of the work proposed and discussed. It also lays the hypothesis for the phenomenon observed and explores different factors that affect double descent, which makes it easy to follow and understand the paper.

VI. Relation to prior work

Prior work in the field of double descent include methods that theoretically analyze the phenomenon in the tractable setting of linear least squares regression and a few others that provide preliminary results for model-wise double descent in convolutional networks trained on CIFAR-10. The work proposed by the paper differs from prior work in a number of ways. First, the paper proposes the Effective Model Complexity (EMC) measure, leading to novel insights like epoch-wise double descent and sample non-monotonicity. Second, the paper also provides an extensive demonstration of double-descent for modern practices including a variety of architectures, datasets, and optimization procedures.

VII. Reproducibility

Enough details are provided in terms of the model architecture, the conditions under which the models were implemented, the datasets, and the factors that contributed to the results that could help reproduce the majority of the work proposed by the authors.

VIII. Additional comments

A stronger theoretical grounding related to the phenomenon of double descent, especially in deep neural networks, would supplement and strengthen the work proposed by the authors.