Paper Write-up: Nakkiran et al. (2019) Deep Double Descent

Group 7

May 15, 2020

One Line Summary

The authors demonstrate that deep learning tasks exhibit "double-descent" wherein model performance gets worse *then better* as a function of model size and generalise this phenomenon to training epochs via their *effective model complexity* metric showing that larger samples sizes can diminsh model performance in some conditions.

Introduction

- Modern neural networks have millions of parameters and yet outperform more parsimonious models in contrast to the notion that *larger models are worse* stemming from the bias-variance trade-off of classical statistics[1]; i.e. models with large numbers of parameters ought to perform worse as they fit noise (i.e. sample variance).
- The best strategy with respect to training time is also questioned: should optimal early stopping - wherein models are trained until an increase in the empirically estimated test error is observed - be employed or do regularised models with zero-training error have better performance?

The authors distinguish *under-parametrised* models from those which are able to *interpolate* data (i.e. achieve approximately zero training error) in the deep learning setting. They claim that the former class of models follow the U-shaped curve of test error following from the bias-variance trade-off but the latter class sees an improvement (i.e. decrease) in test error from increased complexity[2].

[1]The distinction between classical statistics and modern regimes is one asserted by the authors.

[2]The authors point out that similar behavior was previously observed in Manfred Opper (1995) *Statistical mechanics of learning: Generalization. The Handbook of Brain Theory and Neural Networks*, pp. 922-925; Manfred Opper (2001) *Learning to generalize. Frontiers of Life*, 3 part 2, pp. 763-775; Advani and Saxe (2017) High-dimensional dynamics of generalization error in neural networks. arXiv preprint arXiv:1710.03667; Spigler et al. (2018) A jamming transition from under-to over-parametrization affects loss landscape and generalization arXiv

Main Contributions

The main novel contributions of the authors in this paper were as follows:

1. Variously demonstrate double descent in CNNs, ResNets and Transformers[3] using the CIFAR 10, CIFAR 100, IWSLT '14 de-en or WMT '14 en-fr datasets (see Table A from the paper and shown below).

2. Define the *effective model complexity* (EMC) metric: the maximum number of samples on which a training procedure can achieve close to zero training error. The EMC depends on the model complexity but also on the training time (measured in epochs) the procedure is allowed.

3. Demonstrate epoch-wise double descent for models with sufficient complexity (i.e. parameters) to interpolate the data.

4. Demonstrate decrease in model performance (increase in test error) with increasing sample size for models lying in a critical interval of model complexity: "more data hurts".

5. The authors also test the effects of label noise[4] remarking that double descent was observed "most strongly in settings with label noise" and comment that label noise constitutes a proxy for model mis-specification, i.e. inaccurate modelling assumptions.

---

preprint arXiv:1810.09665; and Geiger et a. (2019) Jamming transition as a paradigm to understand the loss landscape of deep neural networks. Physical Review E, 100(1):012115. They note most significantly that the phenomenon was first postulated in generality by Belkin et al. (2018) who named it *double descent* and demonstrated it for decision trees, random features, and 2-layer neural networks using an l_2 loss on a variety of learning tasks including MNIST and CIFAR-10.

[3] See also this Google AI Blog post on transformers.

[4] Corruption of the training sample labels. Here done experimentally by mislabelling the outcome label (the y_i elements of the 2-tuple in the set, S) with some probability, p, i.e. according to a Bernoulli distribution.

A SUMMARY TABLE OF EXPERIMENTAL RESULTS

					Double-Descent		
Dataset	Architecture	Opt.	Aug.	% Noise	Model	Epoch	Figure(s)
CIFAR 10	CNN	SGD	✓	0	✗	✗	5, 27
			✓	10	✓	✓	5, 27, 6
			✓	20	✓	✓	5, 27
				0	✗	✗	5, 25
				10	✓	✓	5
				20	✓	✓	5
		SGD + w.d.	✓	20	✓	✓	21
		Adam		0	✓	–	25
	ResNet	Adam	✓	0	✗	✗	4, 10
			✓	5	✓	–	4
			✓	10	✓	✓	4, 10
			✓	15	✓	✓	4, 2
			✓	20	✓	✓	4, 9, 10
		Various	✓	20	–	✓	16, 17, 18
(subsampled)	CNN	SGD	✓	10	✓	–	11a
		SGD	✓	20	✓	–	11a, 12
(adversarial)	ResNet	SGD		0	Robust err.	–	26
CIFAR 100	ResNet	Adam	✓	0	✓	✗	4, 19, 10
			✓	10	✓	✓	4, 10
			✓	20	✓	✓	4, 10
	CNN	SGD		0	✓	✗	20
IWSLT '14 de-en	Transformer	Adam		0	✓	✗	8, 24
(subsampled)	Transformer	Adam		0	✓	✗	11b, 23
WMT '14 en-fr	Transformer	Adam		0	✓	✗	8, 24

Formalised Hypothesis

Defining a *training procedure*, $\mathcal{T}$ as any procedure taking as input a set $S = \{(X_1, y_1), \ldots, (x_n, y_n)\}$ of labelled training examples and giving as output a classifier $\mathcal{T}(S)$ mapping data to labels, the authors define *effective model complexity* of $\mathcal{T}$ w.r.t. distribution $\mathcal{D}$ and with parameter $\epsilon \geq 0$[ˆchoose-epsilon] to be the maximum number of samples, n, on which $\mathcal{T}$ achieves on average ≈ 0 training error:

$$EMC_{\mathcal{D},\epsilon}(\mathcal{T}) \triangleq max\{n \mid \mathbb{E}_{S\sim D^n}[Error_S(\mathcal{T}(S))] \leq \epsilon\}$$

where $Error_S(M)$ is the *mean error* of model M on training samples S.

They thus propose the following hypotheses:
- **Under-parametrised regimes**: If $EMC_{\mathcal{D},\epsilon}(\mathcal{T})$ is *sufficiently smaller* than the sample size n, perturbations to $\mathcal{T}$ that increase its complexity will **decrease** its test error.
- **Over-parametrised regimes**: If $EMC_{\mathcal{D},\epsilon}(\mathcal{T})$ is *sufficiently larger* than the sample size n, perturbations to $\mathcal{T}$ that increase its complexity will **decrease** its test error.
- **Critically parametrised regimes**: If $EMC_{\mathcal{D},\epsilon}(\mathcal{T}) \approx n$, perturbations to $\mathcal{T}$ that increase its complexity may **increase or decrease** its test error.

Experimental Methods

The authors use:
- **ResNets** - ResNet18s are parametrised by varying the width hyperparameter of convolutional layer, i.e. number of filters: specifically layer widths $(k, 2k, 4k, 8k)$ for various k; default $k = 64$.
- **Standard Convolutional Neural Networks (CNNs)** - 5-layer CNNs with 4 convolutional layers of widths $(k, 2k, 4k, 8k)$ for various k and a fully-connected layer[5].
- **Transformers** - 6-layer encoder-decoder scaling the size of the network by modifying the embedding dimension, d_{model} and setting the width of the fully-connected layers according to $d_{ff} = 4 \cdot d_{model}$
- **Optimisation** - ResNets and CNNs were trained with cross-entropy loss and (1) Adam with learning rate 0.0001 for 4k epochs and (2) Stochastic Gradient Descent (SGD) with learning rate $\propto \frac{1}{\sqrt{T}}$ for 500k gradient steps. Transformers were trained for 80k gradient steps with 10% label smoothing and no drop-out.

Label noise was added experimentally by changing the correct label in a dataset to a uniformly random incorrect one with probability p.

The main experimental result of the paper is empirical validation of this hypothesis for different datasets, model architectures and optimisation methods, as well as different interpolation thresholds gained by varying the number of model parameters, length of training, amount of label noise[6] in the distribution and the number of training samples.

Model-wise Double Descent

The fisrt key state of the learning lifecycle of a model is the model wise double descent phenomenon. This phenomenon occurs when the model is underparameterized. In other words, it's the state when any perturbation of the training procedure that increases the model effective complexity will decrease the test error.
The model wise phenomenon involves the variation of its complexity by varying the parameters of the neural network. To achieve this, variation of the width network is needed, and so increase in the number of filters of the convutionnal layers. In this article, the authors highlight this fact by the increase of the test error around an interpolation threshold. In that case, the peak in test error occurs around the interpolation threshold when the models are just barely large enough to fit the train set.
The model-wise double descent phenomenon also shows that changes which affect

[5]The authors note: "the CNN with width $k = 64$, can reach over 90% test accuracy on CIFAR-10 with data augmentation

[6]Corruption of the training sample labels. Here done experimentally by mislabelling the outcome label (the y_i elements of the 2-tuple in the set, S) with some probability, p, i.e. according to a Bernoulli distribution.

the interpolation threshold (such as changing the optimization algorithm, the number of train samples, or the amount of label noise) also affect the location of the test error peak correspondingly. To illustrate this, the authors has played with different parameters. On the one hand, they have shown how much easier it is to see the double descent phenomenon for higher label noise than for smaller. On the other hand, they have worked with data augmentation, which is a method to improve the performance of the model and which prevents overfitting. They have shown that the width parameter needs to be much larger to descend again to its original point than without the process of increasing the amount of training data.

Epoch-Wise Double Descent

In this section, we will focus on how the training time can be a parameter that can make neural networks undergo the double descent phenomenon.

Firstly, we must explain the behavior of models depending how complex they are :

- **Small size models** (Width Parameter in [1,6]) : These models won't go through double descent. These models will never reach the threshold because they are under-parameterized. The error decreases slowly and never increases. The more the epochs goes on, the more the test error decrease, never reaching zero.

- **Medium size models** (Width Parameter in [7,20]) : Similarly to the previous models, medium models won't experience the double descent, instead they will follow the classical U-shape test error curve. This shape is often associated with the bias-variance trade off. Too much training time make the neural network over-fit the training data. We can overcome this issue by setting an optimal early stopping. The learning will stop as soon as we have reached the minimum error test.

- **Large size models** (Width Parameter in [21,∞]) : In this situation, and this situation only according to this empirical paper, models can go through the double descent phenomenon. When increasing the number of epoch, the test error will first decrease, then will increase near the interpolation threshold, and finally will decrease again.

We usually know that we have to choose between good predictions or good fitting with our data. Since neural networks goal is mainly to make prediction, we prefer the number of epochs where test error is the lowest. This is known as the bias variance trade off.

However, this paper suggests that with large size models, we can surpass this problem if we increase the training time, thanks to the double descent phenomenon.

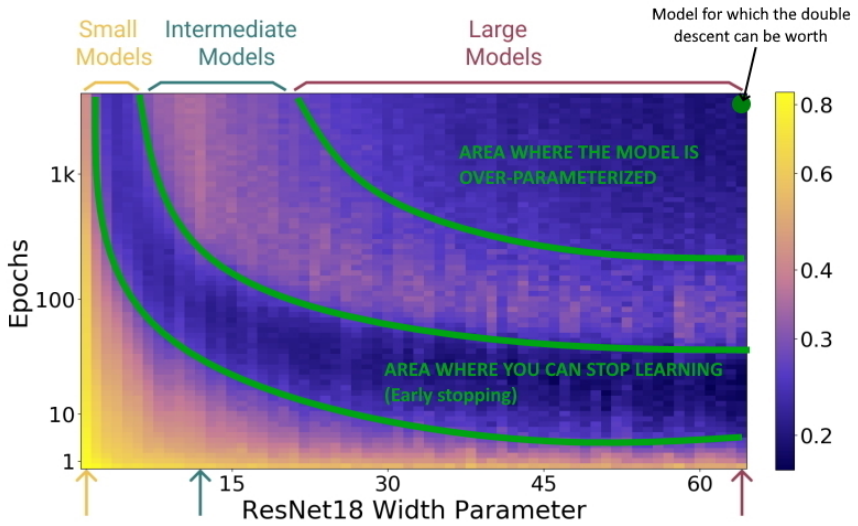

Figure 1: Test error over (Model Size × Epochs)

Sample-Wise Non-Monotonicity

In this section, we will discuss the outcomes and forms of double descent while increasing the size of our training sample.

From most of our classes, we have learned that more data is always better. However, we cannot really say that as we add data points, the test error is going to decrease. Hence the term non-monotonicity. In particular, we notice that as we increase the training sample size, there is a specific point where this increase leads to an increase in the test error. This happens where EMC is approximately equal to our sample size.

Another thing worth mentioning is that the increase in training data points results to a "shift" of the curve to the right, meaning a probable decrease in test error but an increase on the dimension of the parameters needed for the model.

Finally changing the model size and the training sample size leads to some interesting results. In some specific cases, after the interpolation interval, the larger model will give us worse test error than the smaller one, while increasing the training data.

Conclusion

To conclude, the authors provide extensive examples with different data sets, architectures and training procedures as well as a new metric to explain this double descent phenomenon.

Early Stopping We can say that many of these effects do not occur, while optimal early-stopping. However there are setting where even with optimal early stopping we get to see that double descent. The authors leave as an open question the way optimal early stopping and double descent are connected.

Label Noise It is clear from the multiple figures that with large label noise, we get a much more evident double descent. However in most cases, label noise is used as a model mis-specification, to make the model harder to train.

Other Notions of Model Complexity The metrics that the authors created can be seen as an improvement regarding the other metrics like Rademacher complexity and VC dimension. EMC depends also on training procedure as well as model architecture and data distribution. Another feature than we get with EMC is the "epoch-wise" double descent and also the effect that data-augmentation has on our interpolation area.

Critique and Perspective

- As mentioned in Belkin et al. (2018), some optimisation methods such as stochastic gradient descent may be sensitive to the starting parameter values and therefore may not converge to optima with low test set loss, thereby masking the double descent phenomenon.

- Nakkiran and colleagues note themselves that they do not have a principled way of assigning the value of ϵ or defining when $EMC_{\mathcal{D},\epsilon}(\mathcal{T})$ is *sufficiently smaller* or *larger* than the sample size n[7].

- The authors demonstrate double descent for a limited number of datasets and proof of its generality as a phenomenon would require either wider systematic empirical studies on diverse datasets or theoretical grounding.

Resources

- Raw data from the authors' experiments are available for download here

[7] "The width of the critical interval [around the interpolation threshold $EMC_{\mathcal{D},\epsilon}(\mathcal{T}) = n$] depends on both the distribution and the training procedure in ways we do not yet completely understand."