ML journalclub

Deep Double Descent: Where bigger models and more data hurt

Esten Høyland Leonardsen
12.10.22

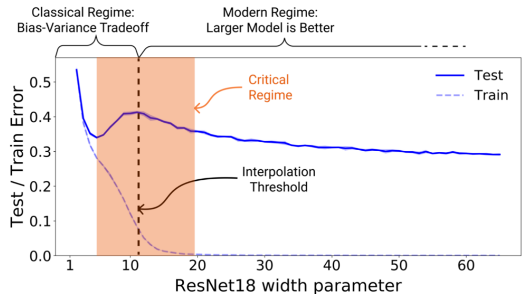
UiO:Life Science, University of Oslo

Introduction

DEEP DOUBLE DESCENT: WHERE BIGGER MODELS AND MORE DATA HURT

Preetum Nakkiran* Harvard University
Gal Kaplun† Harvard University
Yamini Bansal† Harvard University
Tristan Yang Harvard University
Boaz Barak Harvard University
Ilya Sutskever OpenAI

"... a variety of modern deep learning tasks exhibit a double-descent phenomenon where [...] performance first gets worse and then gets better."

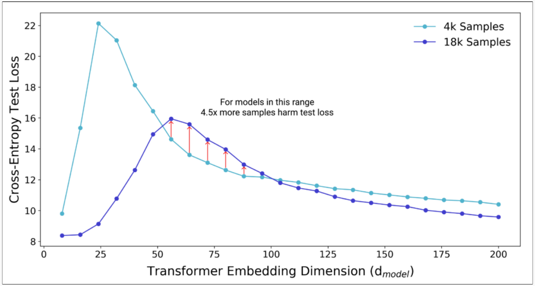

Classical Regime:
Bias-Variance Tradeoff
Modern Regime:
Larger Model is Better
Critical
Regime
Interpolation
Threshold
Test
Train
Test / Train Error
0.0
0.1
0.2
0.3
0.4
0.5
1
10
20
30
40
50
60
ResNet18 width parameter

Introduction

Effective model complexity (EMC): Maximum number of samples on which a model can reach zero training error

- Depends on data distribution, model architecture, and *training procedure* - increasing training time will increase EMC
- Test error peaks around the point where EMC matches the number of samples, increasing the number of samples shifts this peak to the right - in some settings more data is worse (?)

Introduction



Introduction

Hypothesis 1: For any natural data distribution $\mathcal{D}$, neural network-based training procedure $\mathcal{T}$, and small $\epsilon > 0$, if we consider the task of predicting labels based on *n* samples from $\mathcal{D}$ then:

- **Under-parameterized regime**: If $\mathrm{EMC}_{\mathcal{D},\epsilon}(\mathcal{T})$ is sufficiently smaller than *n*, any pertubation of $\mathcal{T}$ that increases its effective complexity will decrease the test error.
- **Over-parameterized regime**: If $\mathrm{EMC}_{\mathcal{D},\epsilon}(\mathcal{T})$ is sufficiently larger than *n*, any perturbation of $\mathcal{T}$ that increases its effective complexity will decrease the test error.
- **Critically parameterized regime**: If $\mathrm{EMC}_{\mathcal{D},\epsilon}(\mathcal{T}) \approx n$, then a perturbation of $\mathcal{T}$ that increases its effective complexity might decrease or increase the test error.

$$
\begin{aligned}
\mathrm{EMC}_{\mathcal{D},\epsilon}(\mathcal{T}) = &\text{maximum number of samples on which a model can} \\
&\text{achieve close to zero training error} \\
= &\text{the number of samples your model can express} \\
= &\text{the expressive power of your model}
\end{aligned}
$$

$$\mathrm{EMC}_{\mathcal{D},\epsilon}(\mathcal{T}) < n = \text{you have more data then your model can express}$$

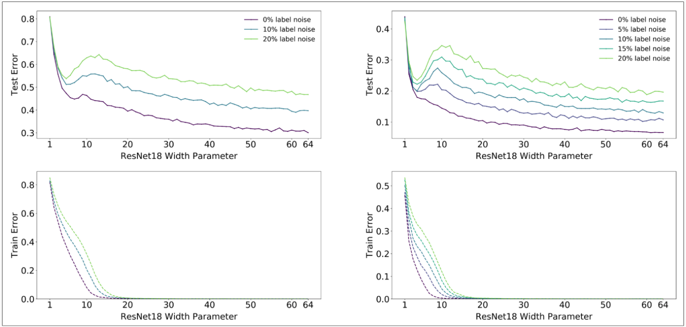
any perturbation of $\mathcal{T}$ that increases its effective complexity =
using a more complex model or training procedure

Hypothesis 1: For any natural data distribution $\mathcal{D}$, neural network-based training procedure $\mathcal{T}$, and small $\epsilon > 0$, if we consider the task of predicting labels based on n samples from $\mathcal{D}$ then:

- Under-parameterized regime: If you have (sufficiently) more data than your model is able to express, a more complex model leads to better performance.
- Over-parameterized regime: If you have (sufficiently) less data than your model is able to express, a more complex model leads to better performance.
- Critically parameterized regime: If you have approximately as much data as your model is able to express, anything could happen

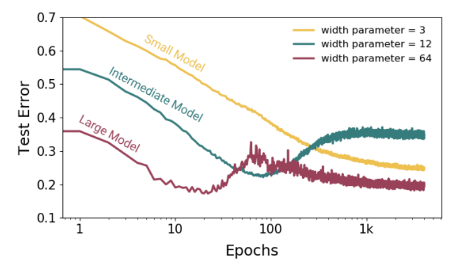
Results

Model-wise double descent



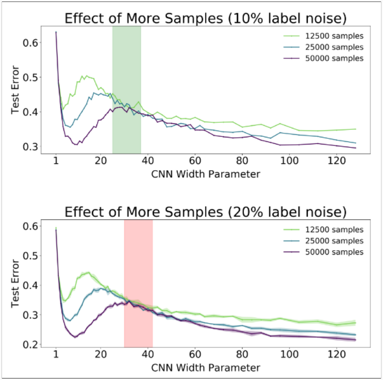
Results

Epoch-wise double descent



Results

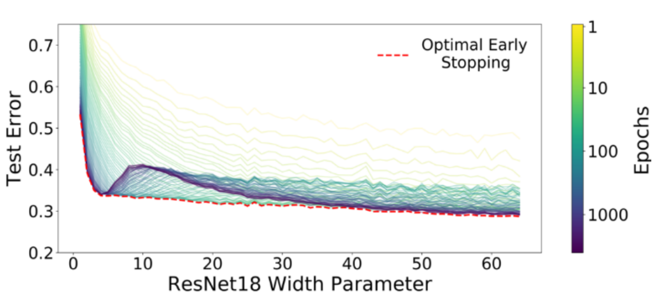
Sample-wise non-monotonicity

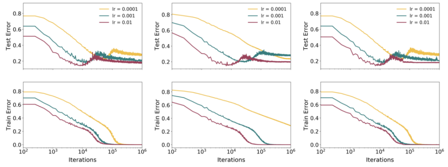


Discussion

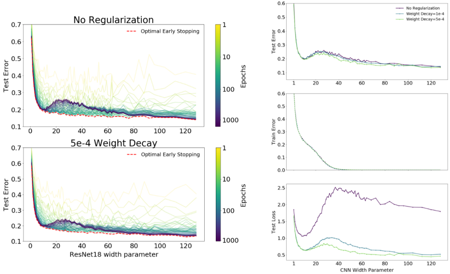
1. Why does this phenomenon occur?
2. Does this matter in practical use cases?
    - Do we observe it? (alternatively, why have I never observed it?)
    - Should we aim for the second descent?

Discussion

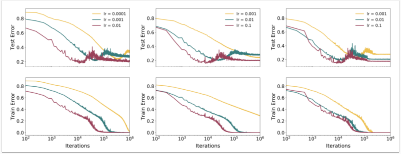

Optimal Early Stopping
Test Error
ResNet18 Width Parameter
Epochs
0.7
0.6
0.5
0.4
0.3
0.2
0
10
20
30
40
50
60
1
10
100
1000



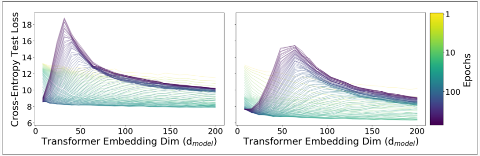
(a) Constant learning rate (b) Inverse-square root learning rate (c) Dynamic learning rate



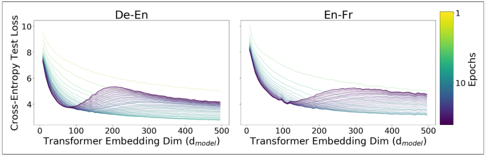
Discussion



lr = 0.0001
lr = 0.001
lr = 0.01
lr = 0.001
lr = 0.01
lr = 0.1
Test Error
Train Error
Iterations

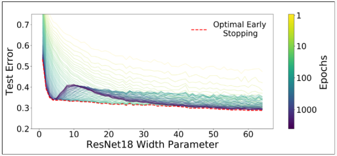


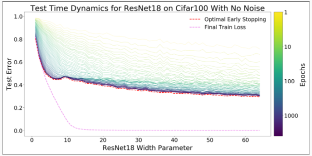
Cross-Entropy Test Loss
Transformer Embedding Dim (d_{model})
Epochs

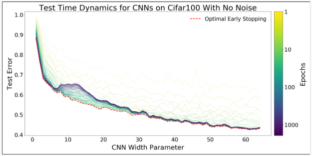


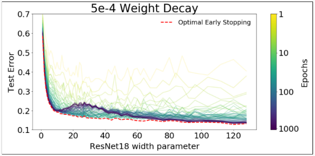
De-En
En-Fr
Cross-Entropy Test Loss
Transformer Embedding Dim (d_{model})
Epochs

Discussion


Optimal Early Stopping
Test Error
ResNet18 Width Parameter
Epochs


Test Time Dynamics for ResNet18 on Cifar100 With No Noise
Optimal Early Stopping
Final Train Loss
Test Error
ResNet18 Width Parameter
Epochs


Test Time Dynamics for CNNs on Cifar100 With No Noise
Optimal Early Stopping
Test Error
CNN Width Parameter
Epochs


5e-4 Weight Decay
Optimal Early Stopping
Test Error
ResNet18 width parameter
Epochs

Discussion

hyperparameter settings decide if/when/how we see the double descent

+

when we see a double descent, the second is not necessarily better than the first

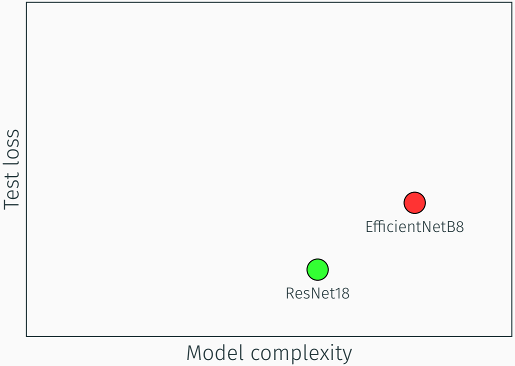
+

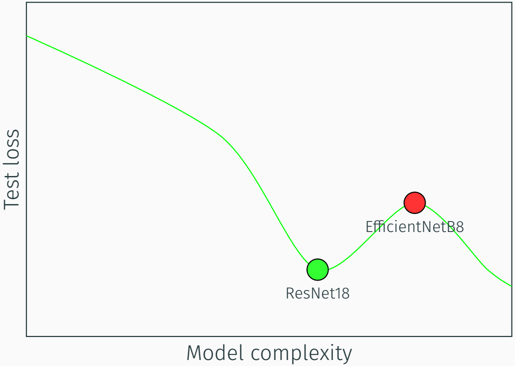
double descents that perform well rely on heavy overparameterization

=

in some cases when training a computationally heavy model for a long time we see a second descent, and some times this outperforms the first descent

Discussion




Test loss
Model complexity
ResNet18
EfficientNetB8

Discussion