Deep Double Descent: Where Bigger Models and More Data Hurt

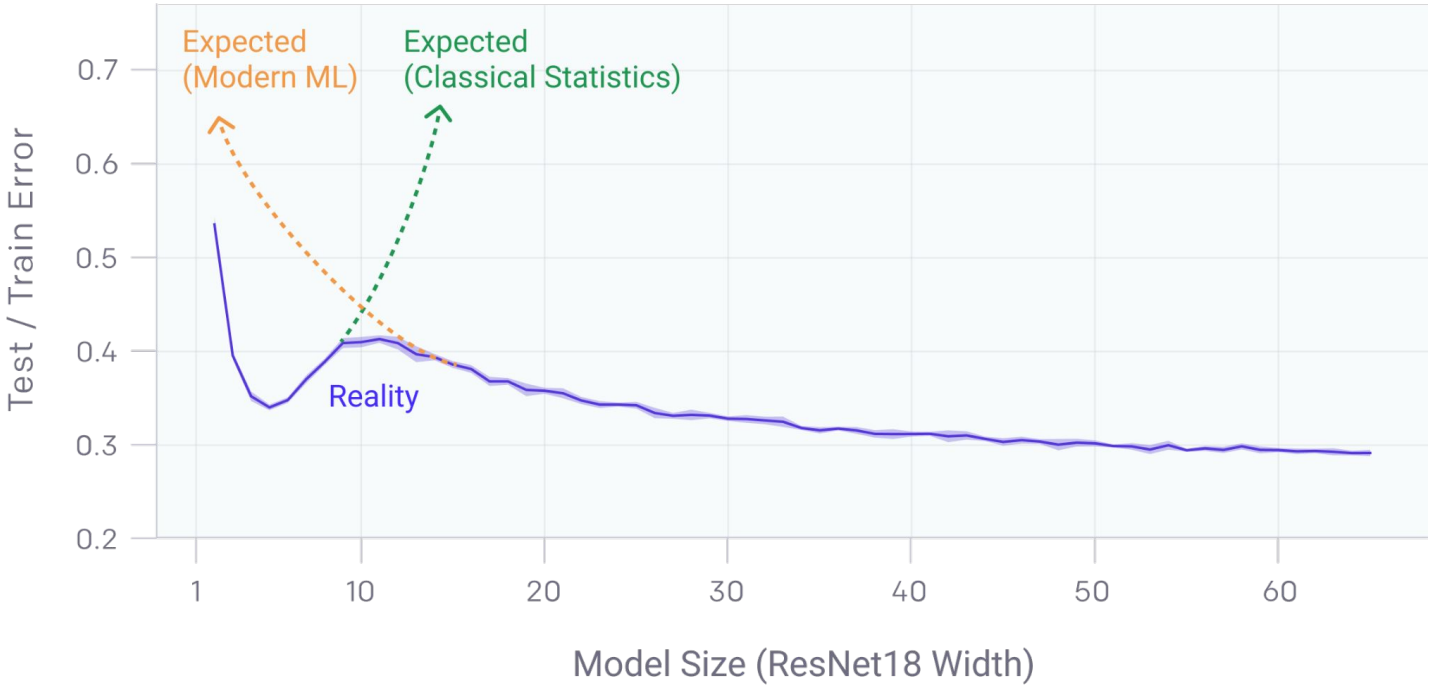
Хамдеева Дилара, Карташев Николай
Даянова Сабина, Григорьев Пётр

1. “Classical wisdoms in DL”
2. Double Descent
3. EMC
   3.1. Определение
   3.2. Гипотеза
4. Эксперименты
   4.1. Model-wise
   4.2. Epoch-wise
   4.3. Sample non-monotonicity
5. Вывод

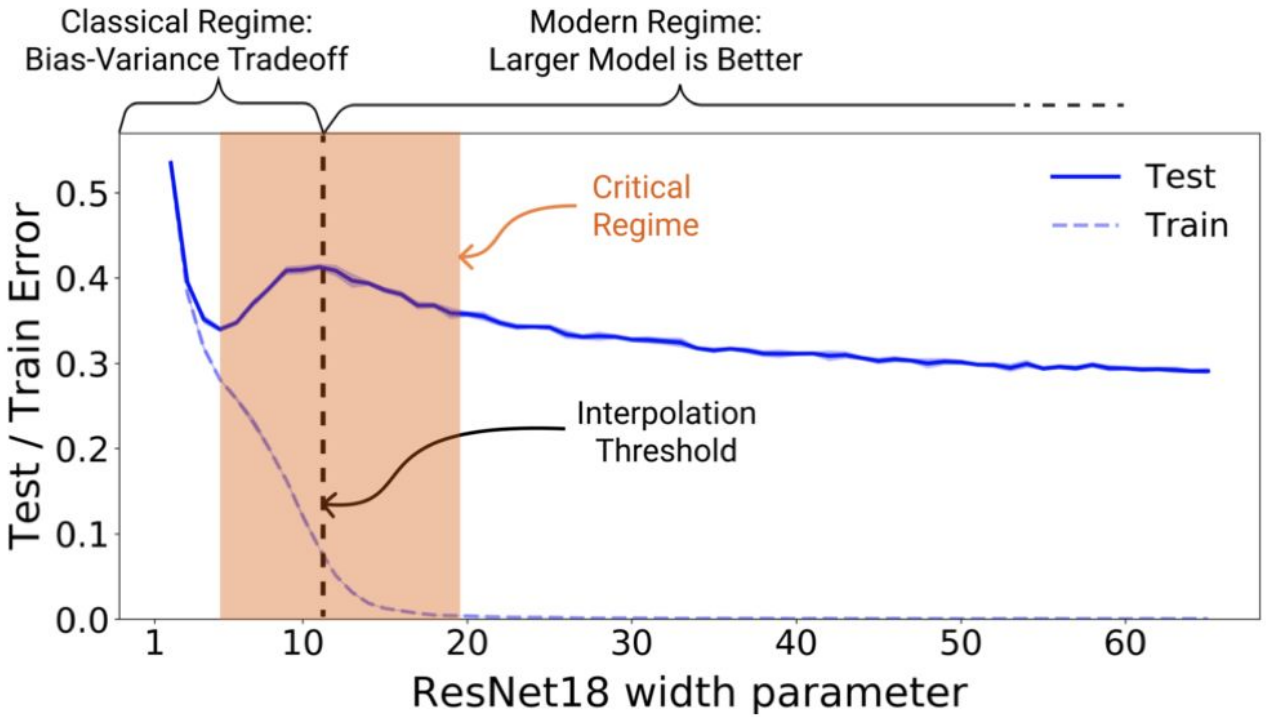
Classical wisdoms in DL

1. Larger models are better
2. More data is always better

Double Descent



Double Descent



EMC (Effective Model Complexity). Определение

Неформально, EMC - максимальное число сэмплов, на которых ошибка на обучении (в среднем) ≈ 0

Definition 1 (Effective Model Complexity) *The* Effective Model Complexity *(EMC) of a training procedure $\mathcal{T}$, with respect to distribution $\mathcal{D}$ and parameter $\epsilon > 0$, is defined as:*

$$\mathrm{EMC}_{\mathcal{D},\epsilon}(\mathcal{T}) := \max\{n \mid \mathbb{E}_{S\sim\mathcal{D}^n}[\mathrm{Error}_S(\mathcal{T}(S))] \leq \epsilon\}$$

where $\mathrm{Error}_S(M)$ *is the mean error of model M on train samples S.*

T - процедура обучения (модель, лосс, аугментации и тд)
S = {(x_1, y_1), … , (x_N, y_N)}
Y_pred = T(S)

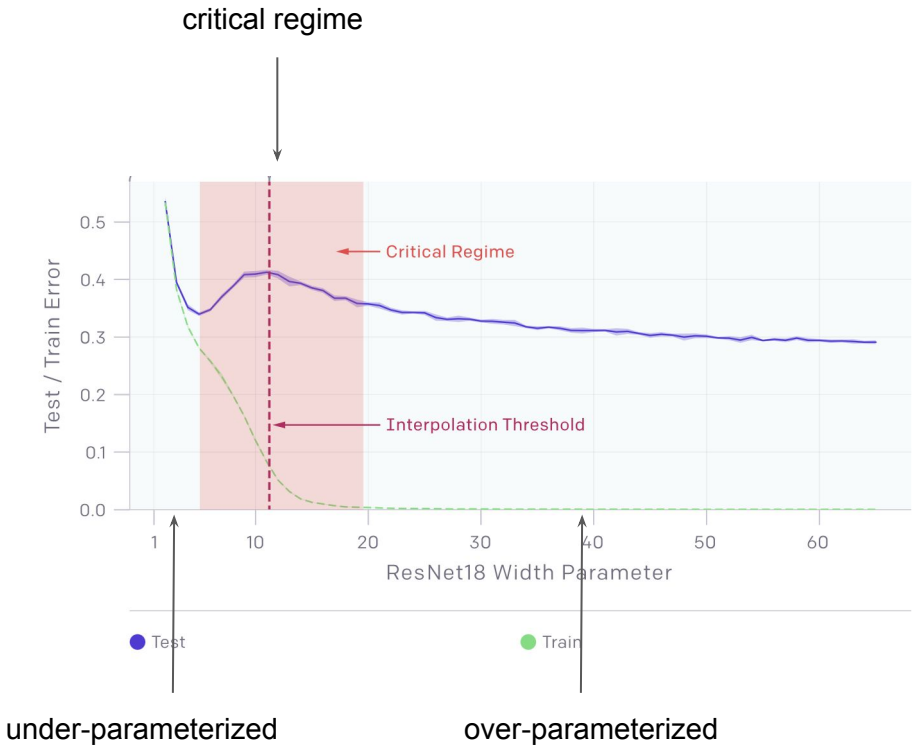
EMC. Гипотеза

Generalized Double Descent hypothesis:

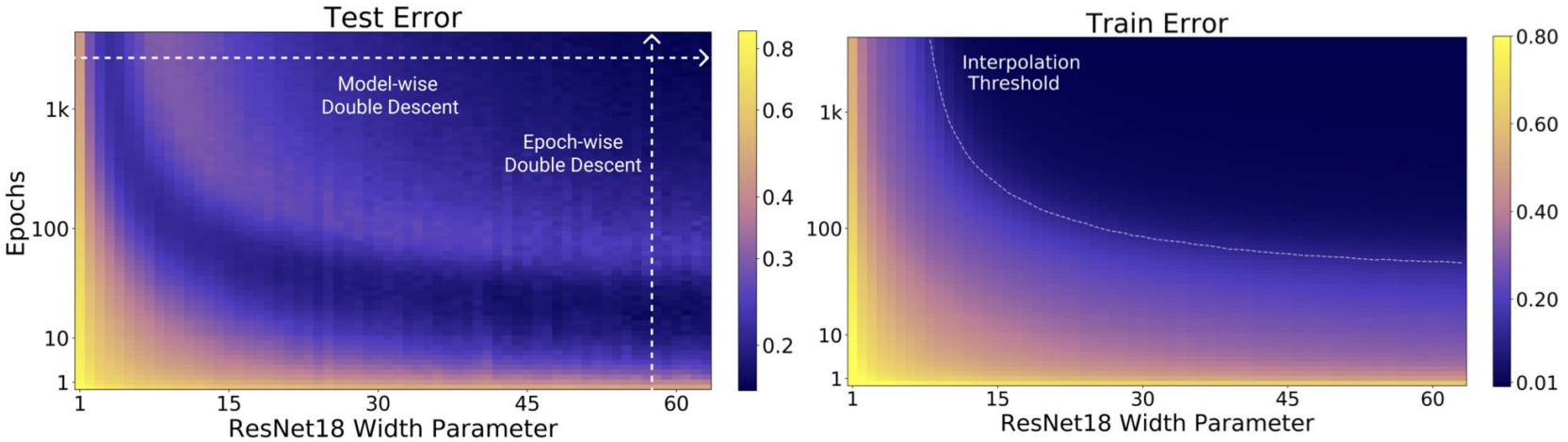
Решаем задачу предсказания меток на N объектах. Тогда

- **Under-parameterized regime**
  - EMC << N
  - увеличение сложности T (процедуры обучения) *уменьшает* ошибку на тесте
- **Over-parameterized regime**
  - EMC >> N
  - увеличение сложности T *уменьшает* ошибку на тесте
- **Critically parameterized regime**
  - EMC ≈ N
  - увеличение сложности T способно либо *уменьшить,* либо *увеличить* ошибку на тесте

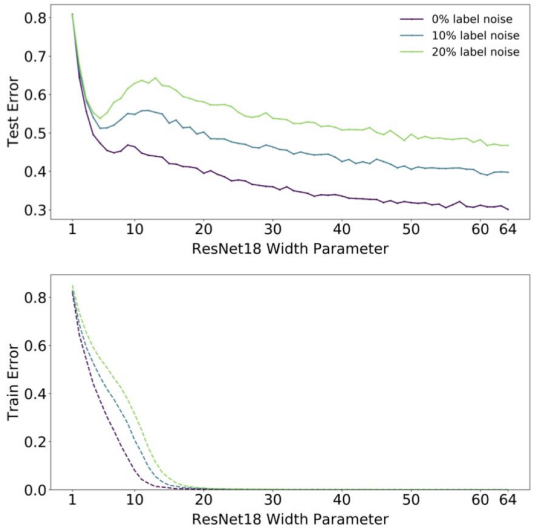
EMC. Regimes


critical regime
Critical Regime
Interpolation Threshold
Test / Train Error
0.5
0.4
0.3
0.2
0.1
0.0
1
10
20
30
40
50
60
ResNet18 Width Parameter
Test
Train
under-parameterized
over-parameterized

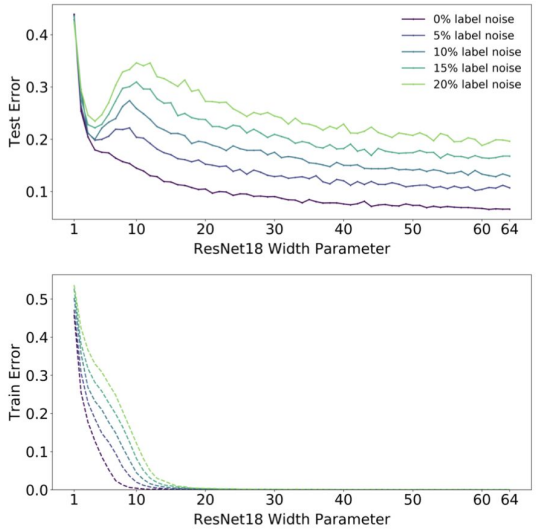
EMC. Regimes


Test Error
Model-wise
Double Descent
Epoch-wise
Double Descent
Epochs
1k
100
10
1
1
15
30
45
60
ResNet18 Width Parameter
0.8
0.6
0.4
0.3
0.2
Train Error
Interpolation
Threshold
1k
100
10
1
1
15
30
45
60
ResNet18 Width Parameter
0.80
0.60
0.40
0.20
0.01

Experiments. Model-wise DD



(a) **CIFAR-100.** There is a peak in test error even with no label noise.

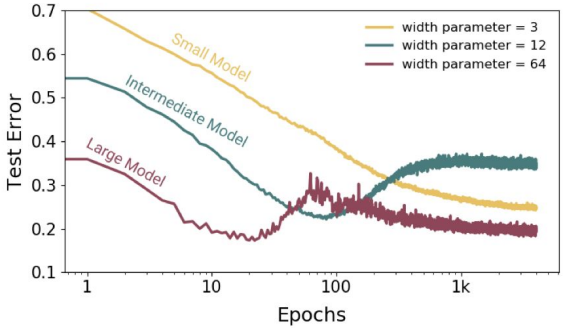


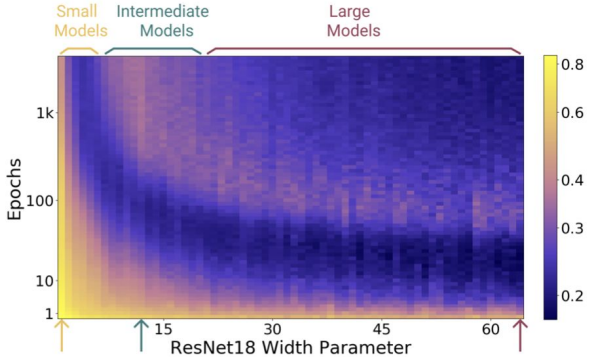
(b) **CIFAR-10.** There is a "plateau" in test error around the interpolation point with no label noise, which develops into a peak for added label noise.

- Пиковые области ярко выражены на зашумленных данных
- EMC находится в пиковой области

Вывод: большие модели не всегда хороши

Experiments. Epoch-wise DD



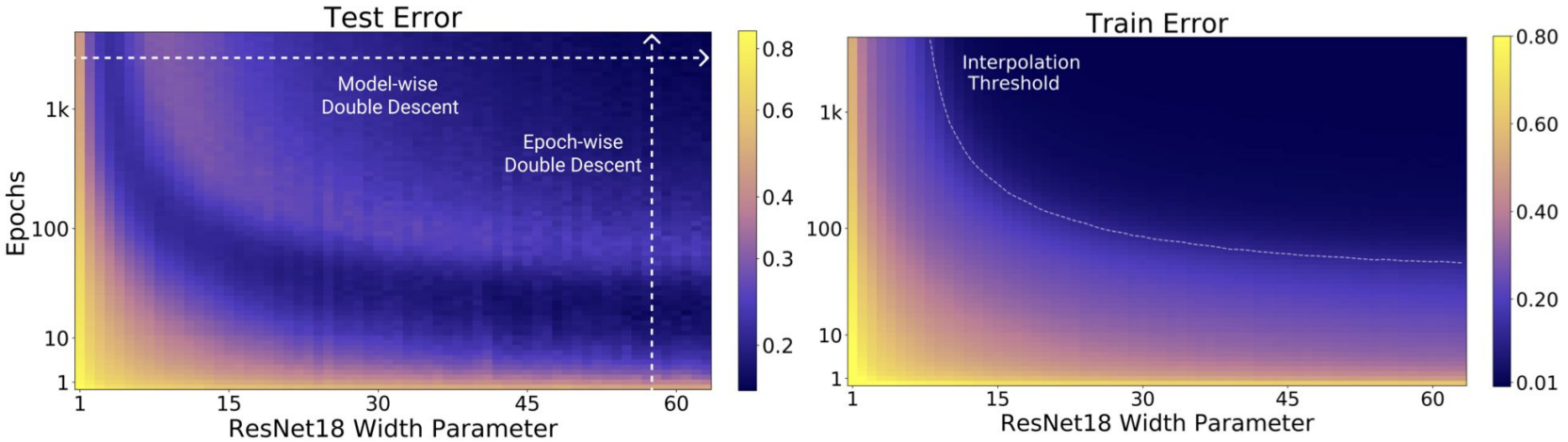


ResNet18 on CIFAR10 with 20% Label noise

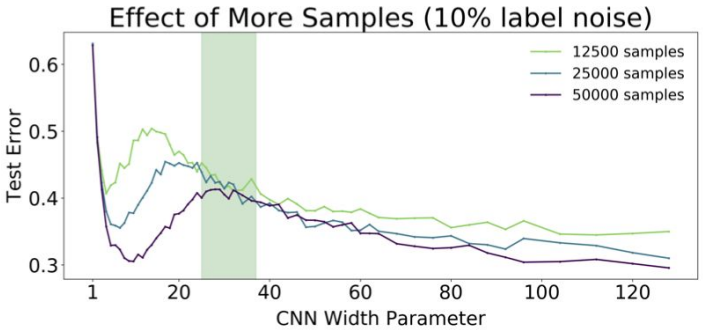
- Большие модели переходят от under- к over-parameterized режиму, (т.е. наблюдается DD) => обучаем дальше
- Средние модели застревают в критическом режиме (плато) => early stopping
- Маленькие модели остаются в under-parameterized режиме, ошибка монотонно убывает => обучаем дальше

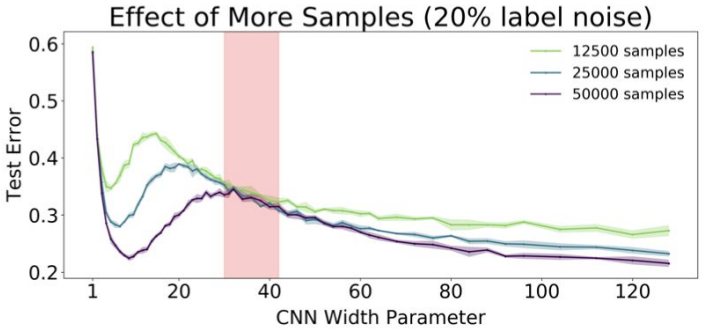
Вывод: более длительное обучение помогает (в ряде случаев)

Model-wise & Epoch-wise DD



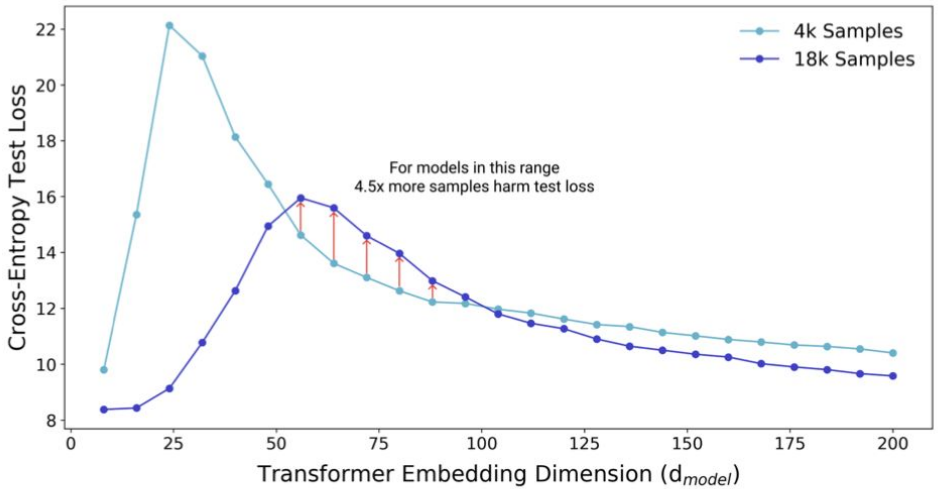
Experiments. Sample non-monotonicity





- Сдвиг пика вправо
- Разница между мин test error в under-parameterized и в over-parameterized уменьшается и становится незначительной.

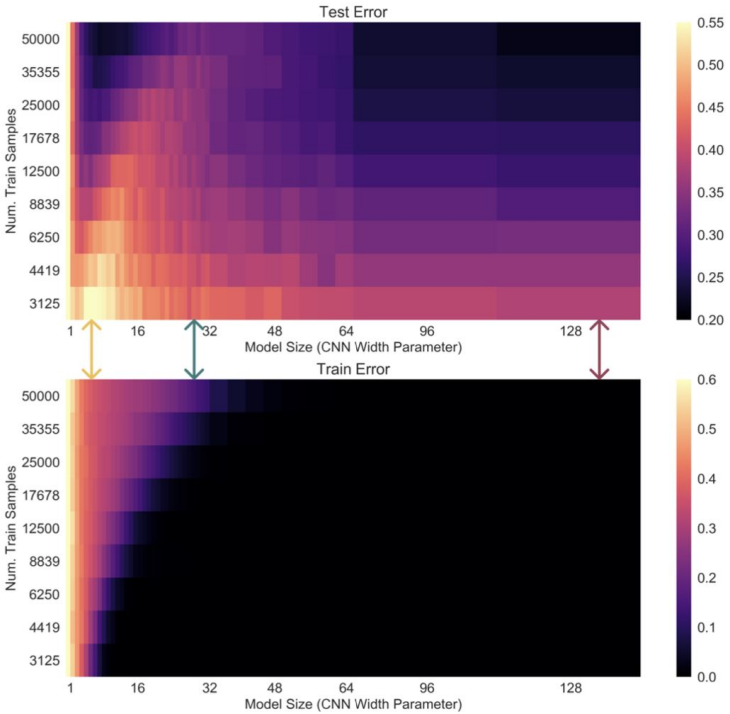
Experiments. Sample non-monotonicity

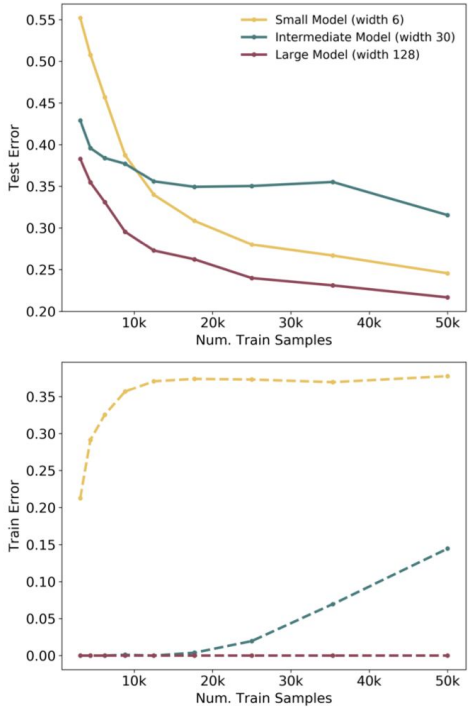


- Сдвиг пика вправо
- В критической области (красным) модели на 18к сэмплах хуже, чем на 4к

Вывод: more data can hurt

Experiments. Sample non-monotonicity





Вывод

- DD зависит не только от сложности модели (model-wise dd), но и от размера датасета (sample non-monotonicity) и от времени обучения (epoch-wise dd)
- more data *can* hurt
- bigger models *can* hurt

Рецензент

Вклад

- Исследуют известный феномен deep double descent
- Проводят очень много экспериментов (статья размером 24 страницы)
- Вводят понятие EMC
- Показывают связь EMC и deep double descent
- Вводят epoch-wise double descent

Плюсы

- Очень много экспериментов
- Интересные объяснения полученных результатов - кратко:

Вокруг критической точки модель может переобучиться чтобы подходить под тренировочный сет, но есть только одна такая модель, поэтому она зависит от каждого шума в данных.

Если модель сложнее, есть много подходящих под train данные моделей, и выбирается наиболее хорошо интерполирующая.

Минусы

- Deep Double Descent не новое понятие
- Переведенную мной цитату слева авторы не доказывают, хотя мне она не кажется следующей из проведенных экспериментов.
- Сложность модели описывается только размером скрытого слоя. Например, непонятно насколько меняется сложность с увеличением размера и добавлением регуляризации
- Очень неформально введена EMC, оформление намного более математическое, хотя на самом деле это просто версия цитаты слева.

Плюсы и минусы по критериям:

- Корректность - утверждений не делается, поэтому все корректно, все попытки объяснить результаты позиционированы как гипотезы

- Значимость - проводится много экспериментов, достаточно чтобы показать что феномен глобально существует в Deep Learning

- Актуальность - достаточно высокая, и даже не только в теории DL.

Выводы

- Насколько красиво написана статья - отлично
- Воспроизводимость - полная, есть код на Pytorch
- Сильные эксперименты, слабая база для выводов. Однако, непонятно как можно было это экспериментально обосновать лучше.
- Я бы хотел увидеть эксперименты с SWA, SAM, исследования поверхности функции потерь, и какой-то промежуточный шаг между самим феноменом и выводом.
- Оценка 7, уверенность 3.5

Практик-исследователь

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

- Статья написана в декабре 2019
- Представлена на ICLR2020, формат – постер.
- Более 340 цитирований

Preetum Nakkiran

Gal Kaplun

Yamini Bansal

- На момент написания статьи получал PhD в Harvard University.
- Сооснователь ML Foundations Group at Harvard.
- ✨Generalization in ML✨
- Сейчас – постдок в UCSD под руководством Михаила Белкина.

Preetum Nakkiran

Gal Kaplun

- PhD in CS at Harvard.
- ✨Generalization in ML✨, ✨Robust Optimization✨

[Submitted on 28 May 2019]

SGD on Neural Networks Learns Functions of Increasing Complexity

Preetum Nakkiran, Gal Kaplun, Dimitris Kalimeris, Tristan Yang, Benjamin L. Edelman, Fred Zhang, Boaz Barak

Yamini Bansal

- PhD in CS at Harvard.
- ✨Generalization in ML✨, ✨Representations✨

[Submitted on 17 Sep 2020 (v1), last revised 15 Oct 2020 (this version, v2)]

Distributional Generalization: A New Kind of Generalization

Preetum Nakkiran, Yamini Bansal

References

[Submitted on 28 Dec 2018 (v1), last revised 10 Sep 2019 (this version, v2)]

Reconciling modern machine learning practice and the bias-variance trade-off

Mikhail Belkin, Daniel Hsu, Siyuan Ma, Soumik Mandal

"Very rich models such as neural networks are trained to exactly fit the data. Such models would be considered over-fit, and yet they often obtain high accuracy on test data."

[Submitted on 18 Mar 2019 (v1), last revised 10 Oct 2020 (this version, v2)]

Two models of double descent for weak features

Mikhail Belkin, Daniel Hsu, Ji Xu

"This article provides a precise mathematical analysis for the shape of an error curve in two simple data models."

Отличия статьи

- Расширяют понятие double-descent. Помимо model-wise DD, вводят такие термины как epoch-wise DD, sample-wise non-monotonicity.
- Проводят обширный анализ этих явлений на разных данных, моделях, методах оптимизации.

Продолжения

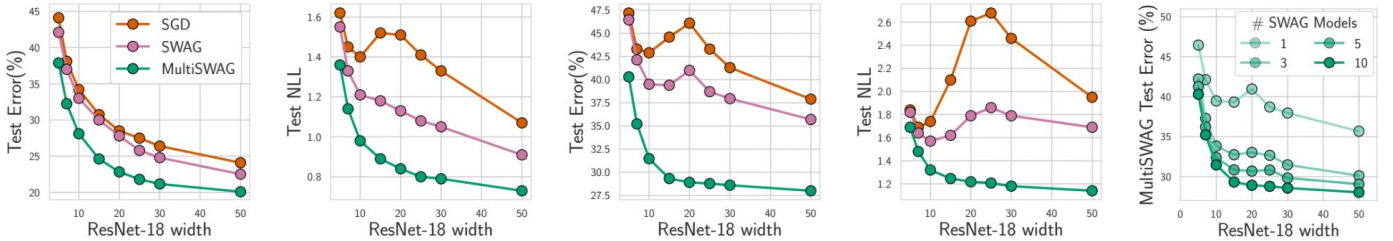
More Data Can Hurt for Linear Regression: Sample-wise Double Descent

Preetum Nakkiran
Harvard University

Bayesian Deep Learning and a Probabilistic Perspective of Generalization

Andrew Gordon Wilson
New York University

Pavel Izmailov
New York University



"We show that Bayesian model averaging alleviates double descent, resulting in monotonic performance improvements with increased flexibility"

Дополнительные исследования и применения

- Дополнительные исследования: такой же анализ (model-wise, epoch-wise DD; sample-wise non-monotonicity) в классических методах ML, генеративных моделях
- Применения: использование моделей в индустрии. Перенос игрушечных примеров малых размеров на реальные масштабы. Зависимость ошибки от роста параметров модели, данных.