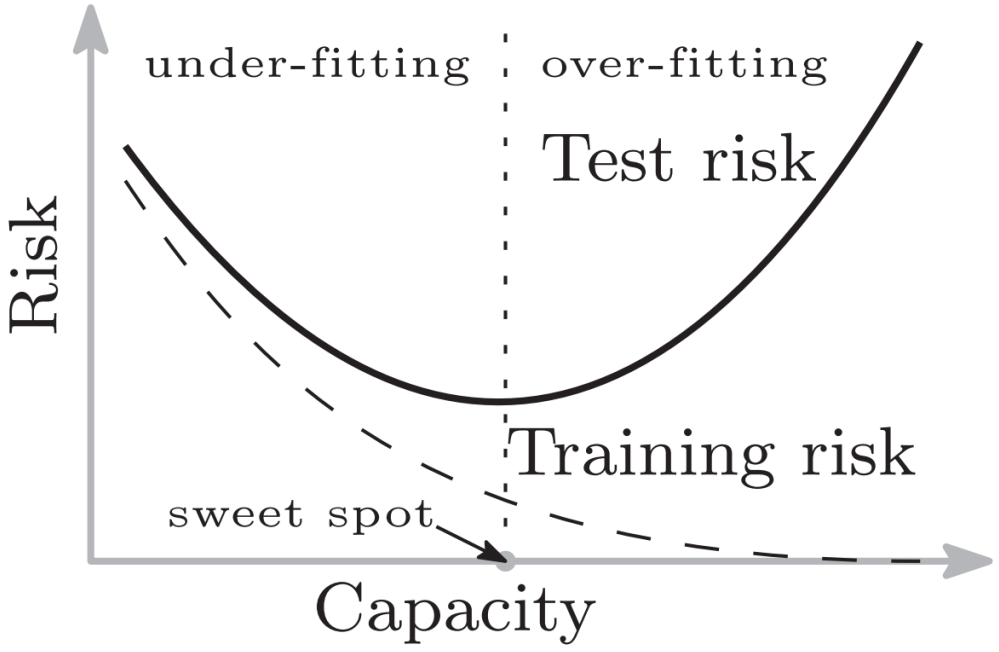
Double Descent

Гальцев Даниил
НИУ ВШЭ
10.04.2020

Классический режим обучения

- Из разложения на смещение и разброс ожидается U-образная зависимость функции потерь от сложности модели

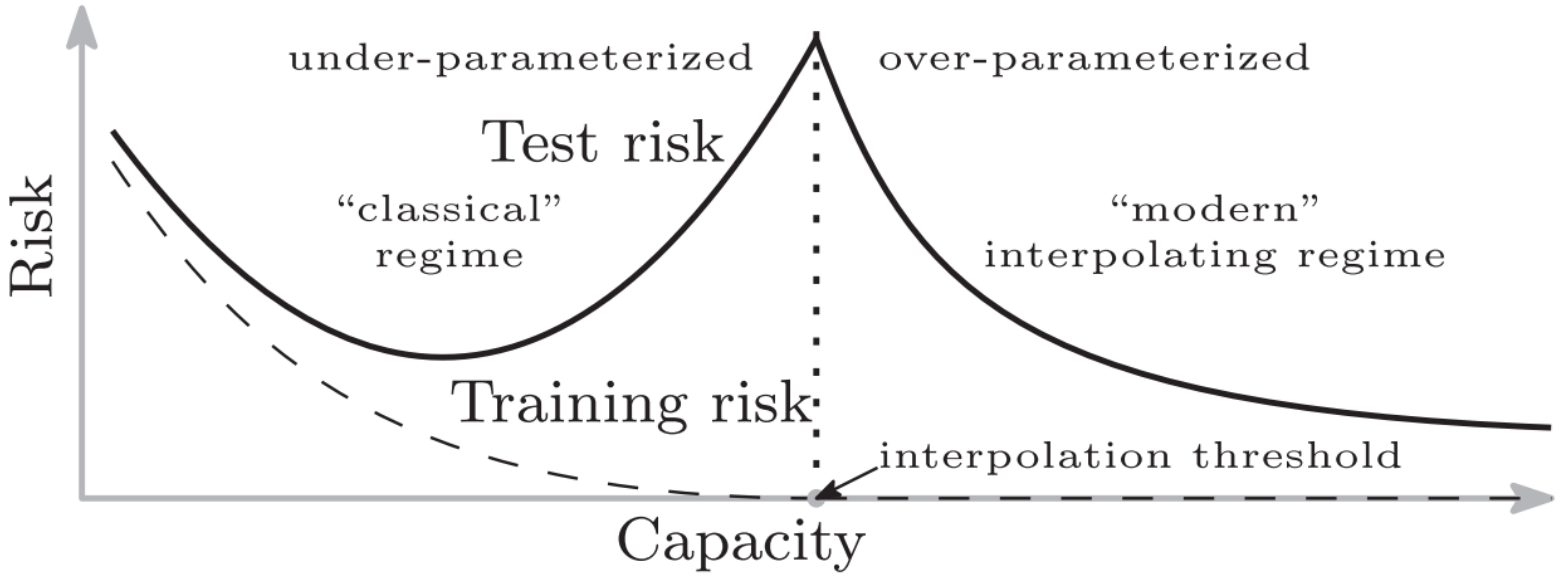


Современный режим обучения

- На практике большие нейронные сети обучают до нулевой ошибки
- Полученные сети показывают хороший результат на тестовой выборке

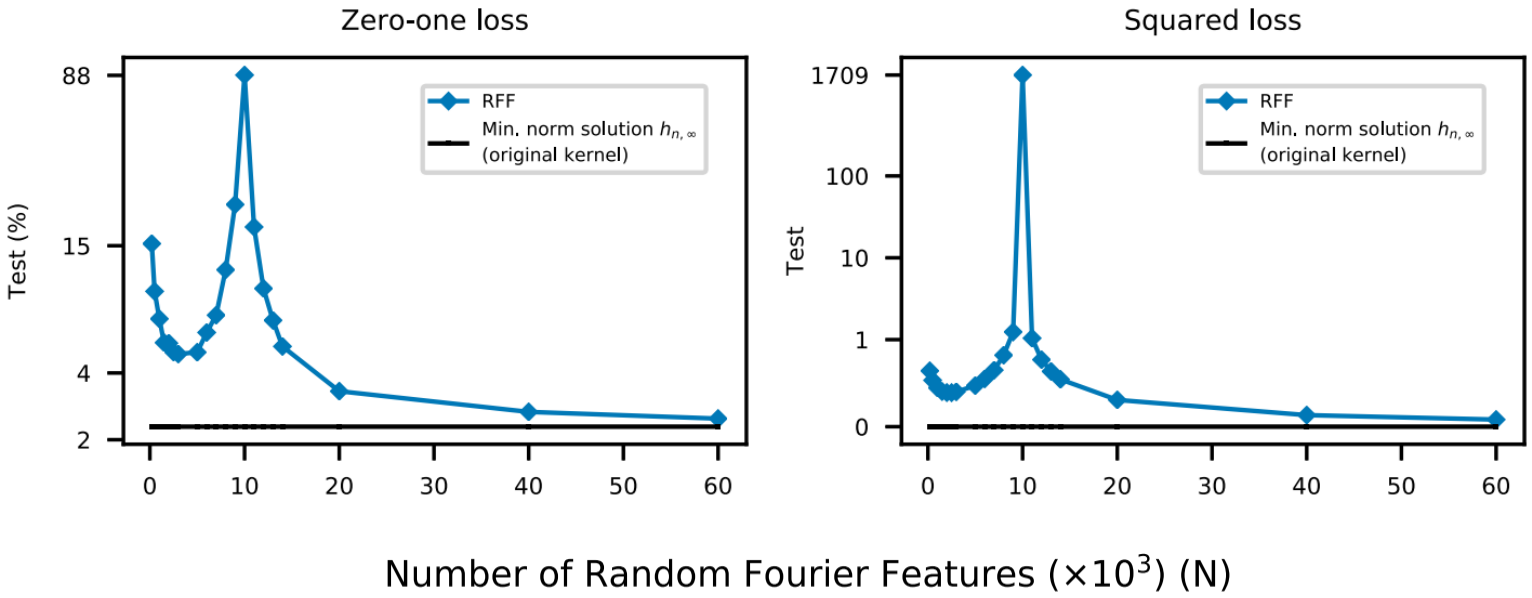
Double Descent

- При увеличении сложности модели происходит двойной спуск функции потерь
- Классический режим заканчивается при интерполяции обучающей выборки
- После этого начинается “современный” режим обучения



Double Descent

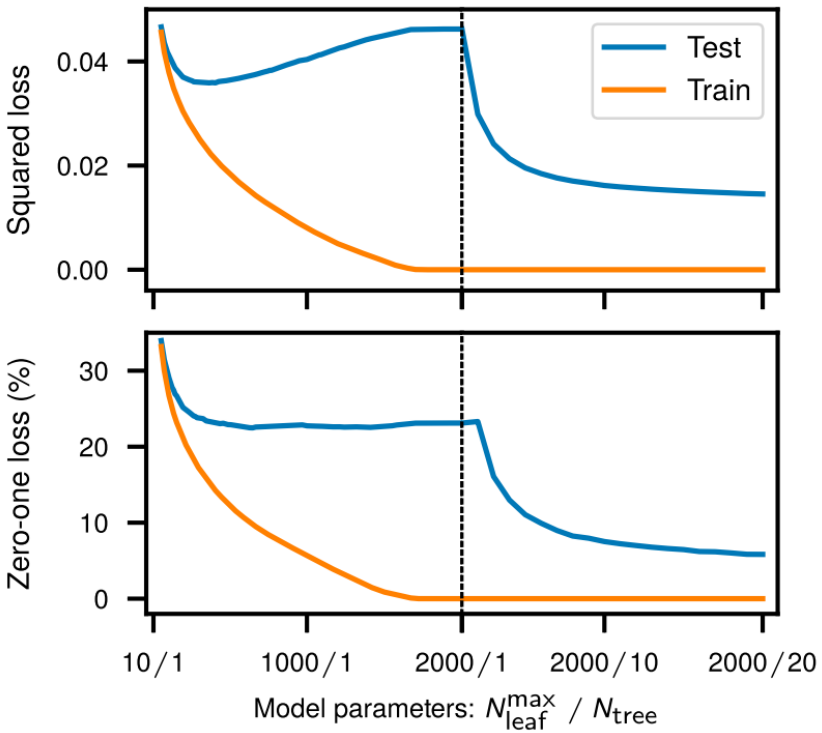
- Этот эффект у различных моделей проявляется по-разному



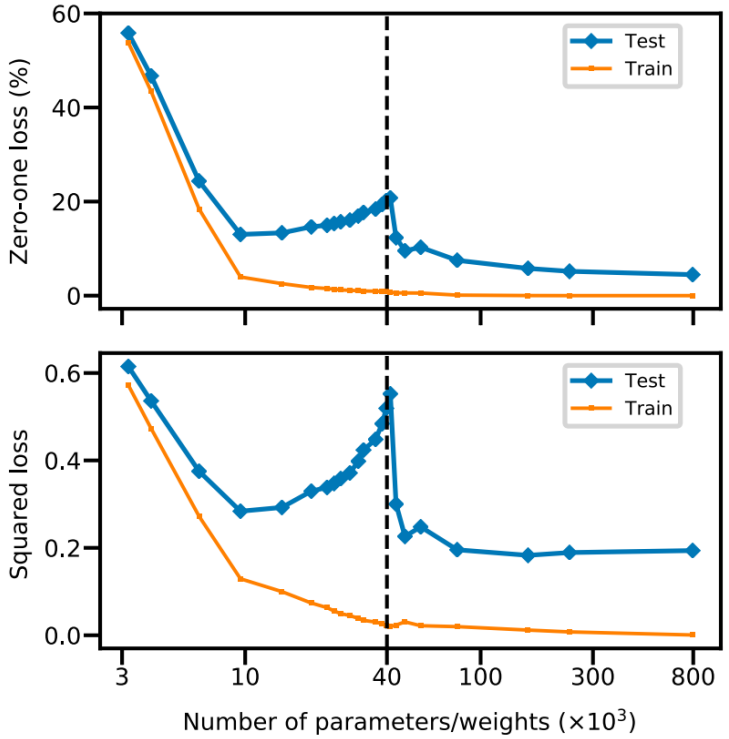
Double Descent

- Этот эффект у различных моделей проявляется по-разному

Случайный лес



Двухслойная нейронная сеть



Effective Model Complexity

EMC – это максимальное количество объектов в выборке, при котором тренировочная процедура в среднем достигает практически нулевой тренировочной ошибки

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

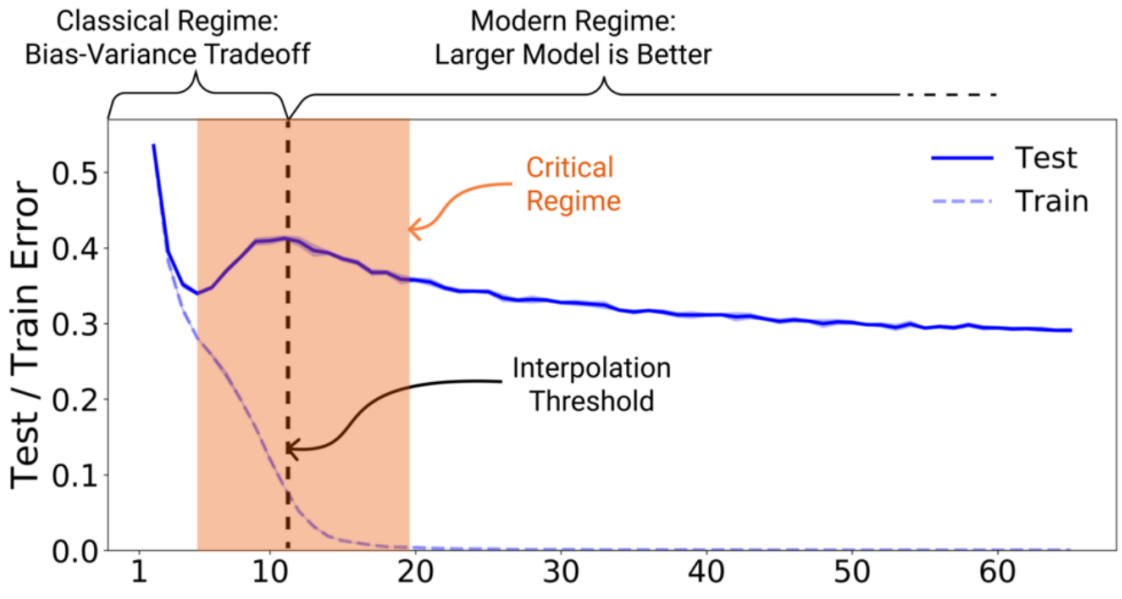
Generalized Double Descent Hypothesis

Гипотеза формулируется для нейронных сетей и заданной выборки размера n:

Under-paremeterized regime. *If* $\mathrm{EMC}_{\mathcal{D},\epsilon}(\mathcal{T})$ *is sufficiently smaller than* n*, any perturbation 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.*

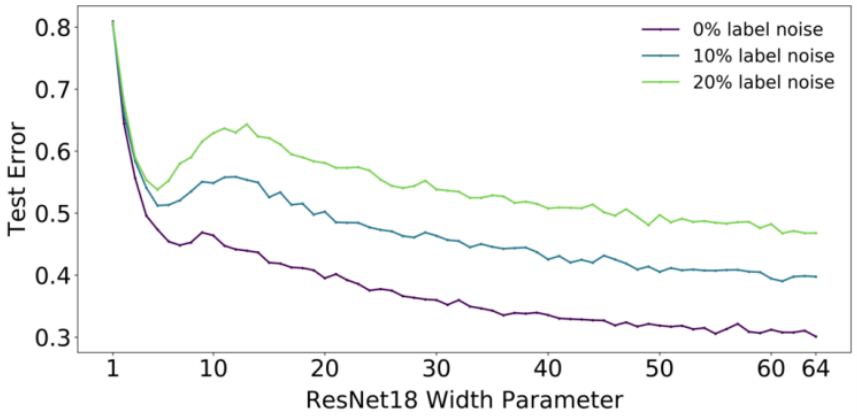
Critically parameterized regime. *If* $\mathrm{EMC}_{\mathcal{D},\epsilon}(\mathcal{T}) \approx n$*, then a perturbation of* $\mathcal{T}$ *that increases its effective complexity might decrease* **or increase** *the test error.*

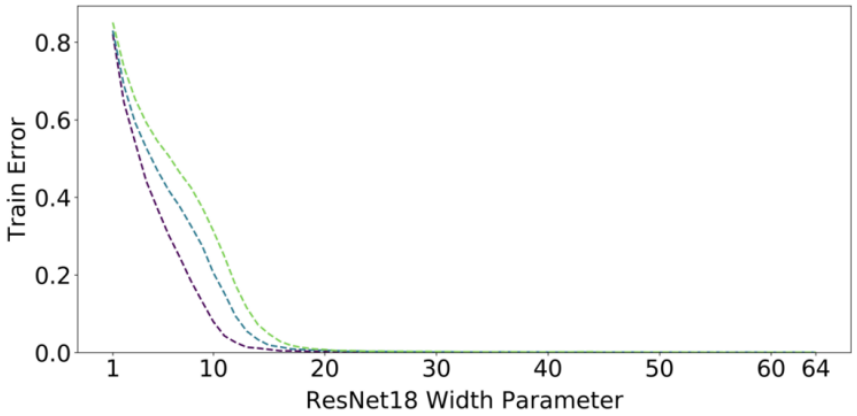


Model-wise Double Descent

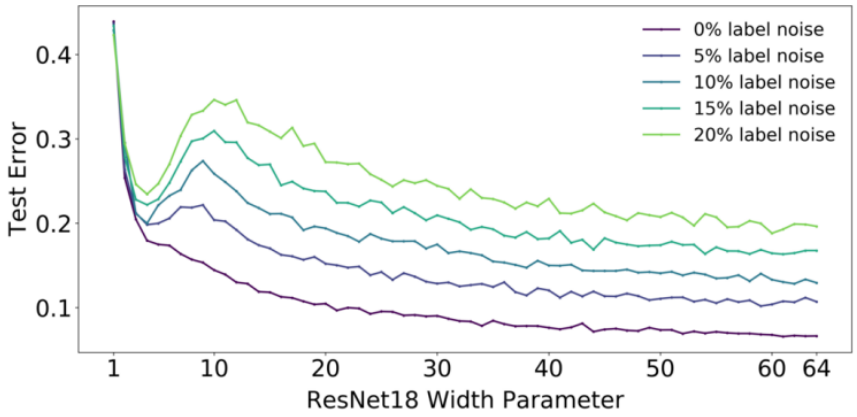
- При увеличении размера модели увеличивается EMC
- При этом наблюдается Double Descent

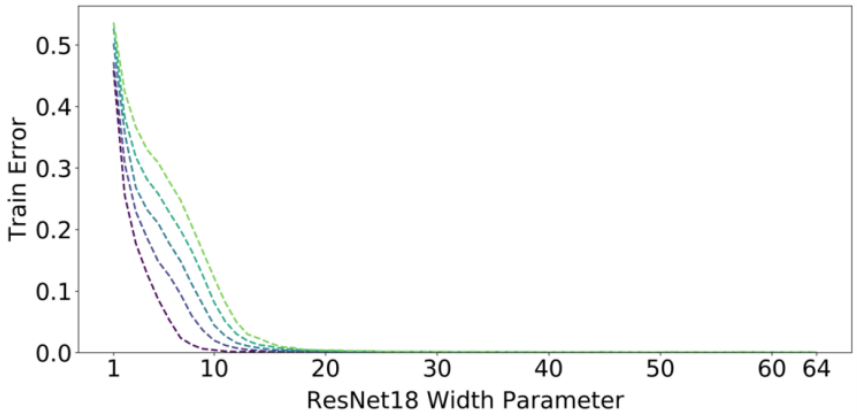
CIFAR-100





CIFAR-10

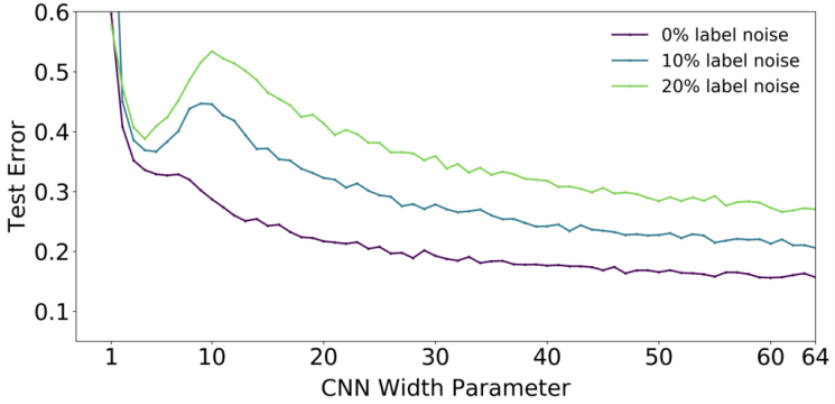


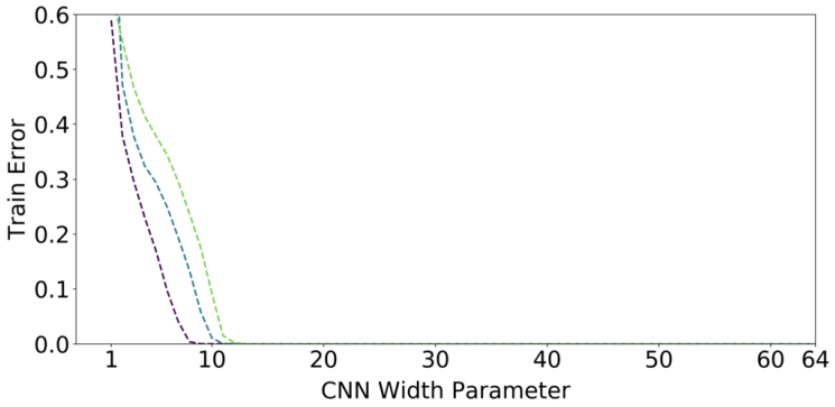


Model-wise Double Descent

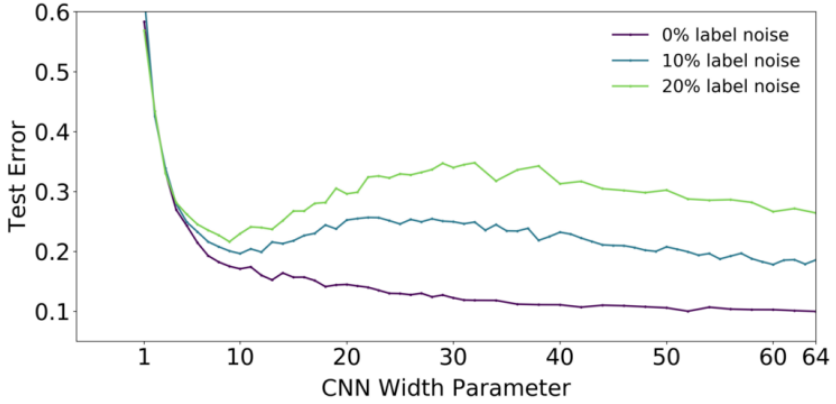
- Аугментация данных смещает порог интерполяции в сторону больших моделей

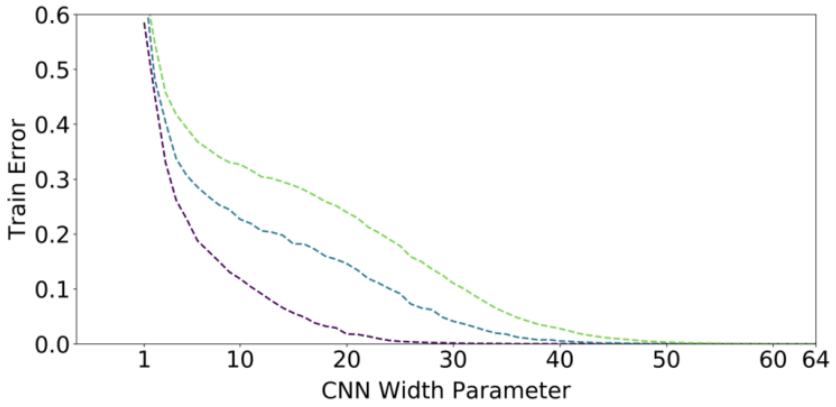
Without data augmentation





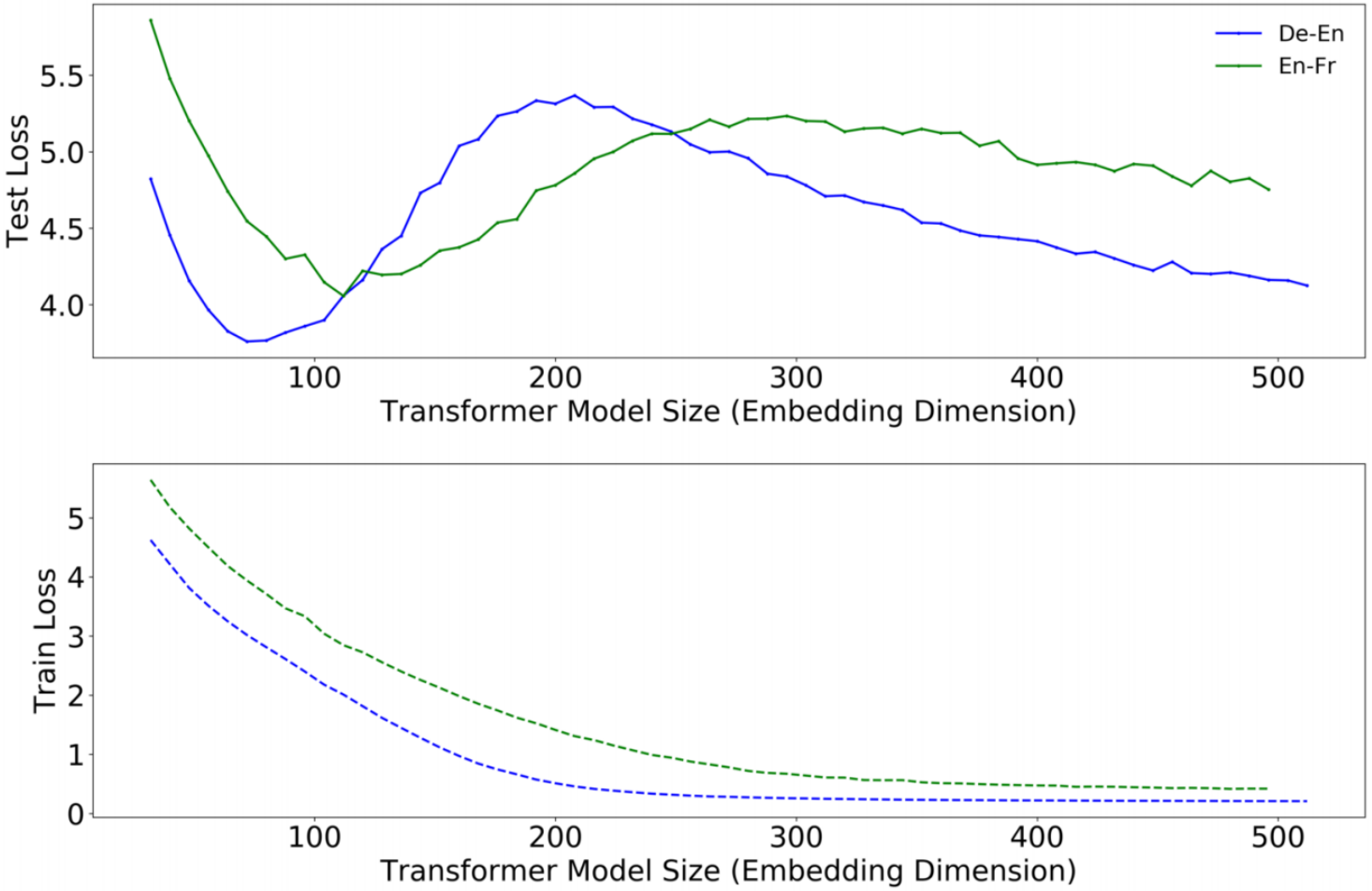
With data augmentation





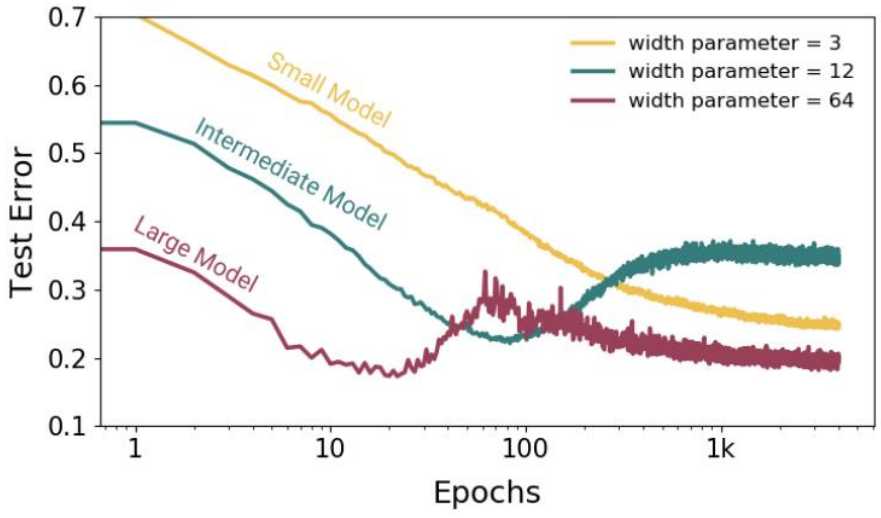
Model-wise Double Descent

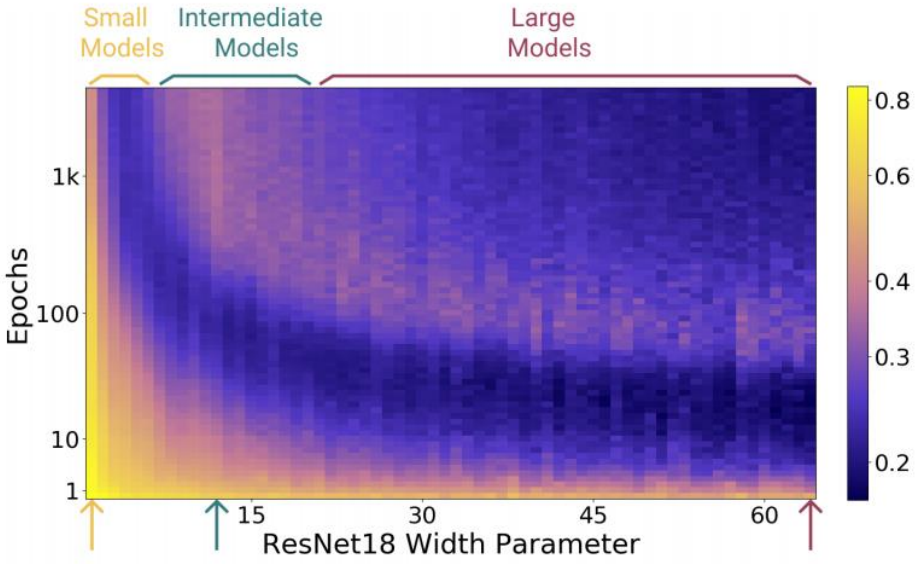
- Double Descent от размера модели наблюдается и для трансформеров



Epoch-wise Double Descent

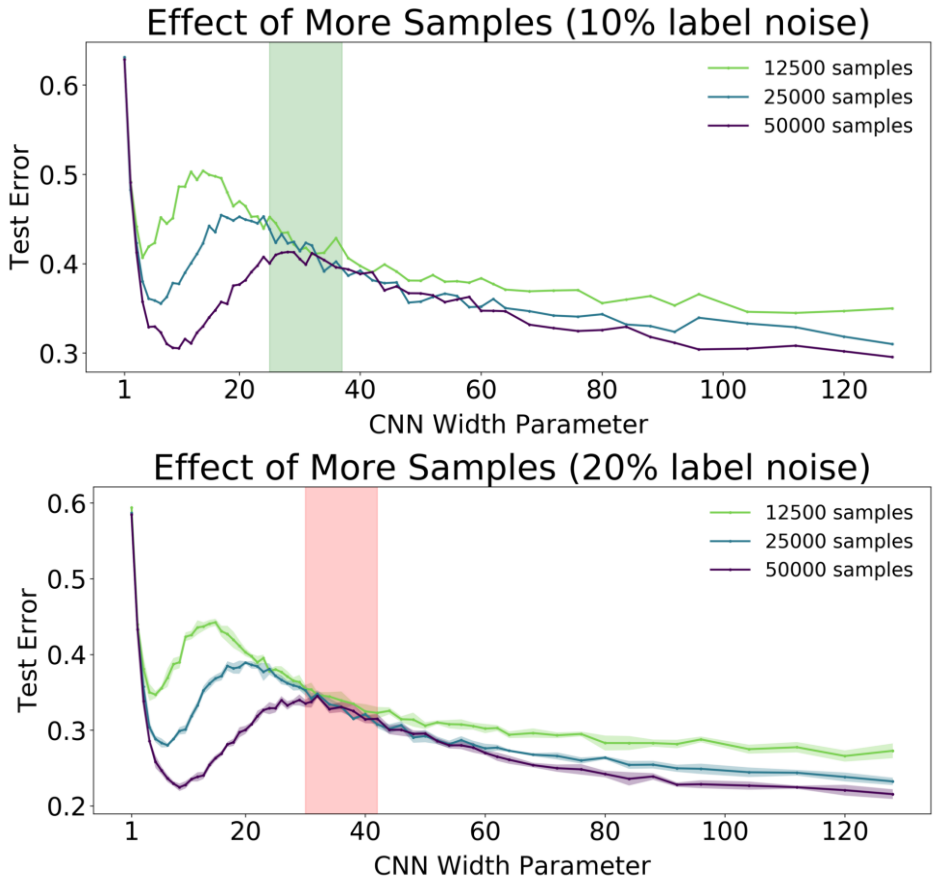
- При увеличении времени обучения увеличивается EMC
- В зависимости от размера модели может наблюдаться Double Descent





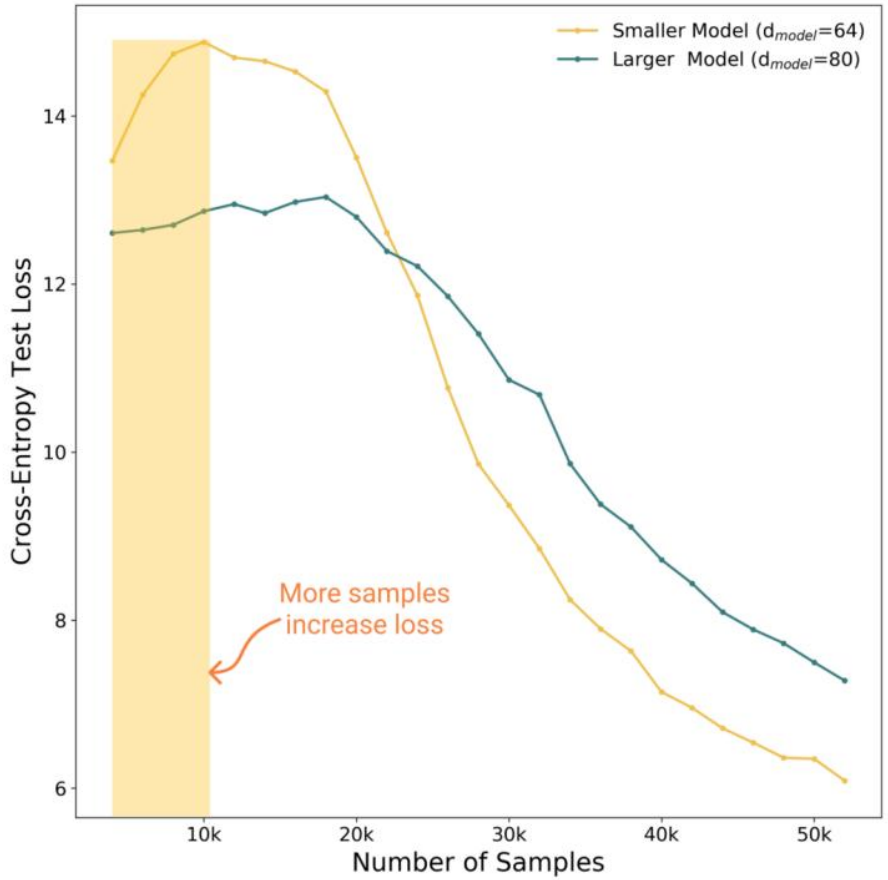
Sample-wise Double Descent

- При изменении размера выборки модель может перейти в другой режим обучения
- Из-за этого наблюдается Double Descent
- При этом при увеличении выборки уменьшается площадь под кривой и порог интерполяции смещается вправо



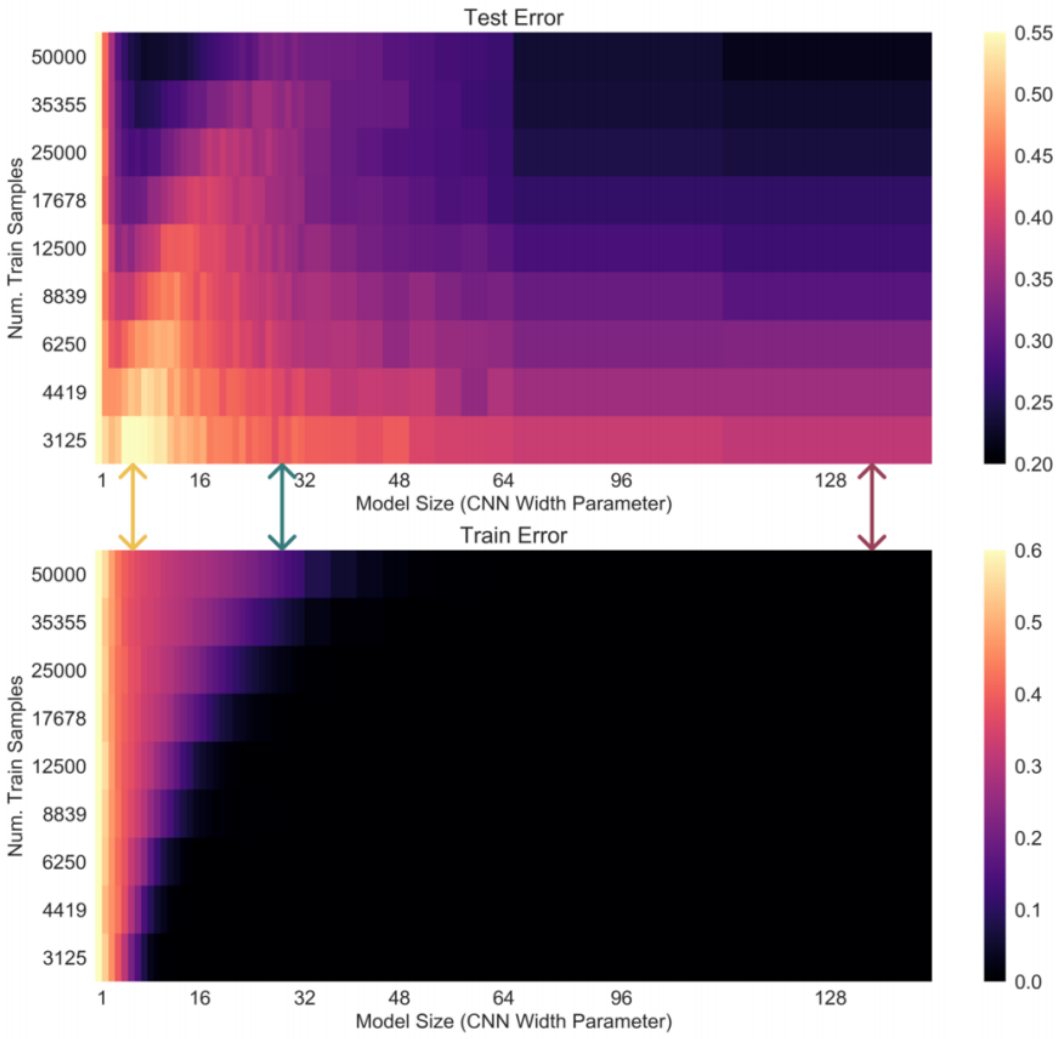
Sample-wise Double Descent

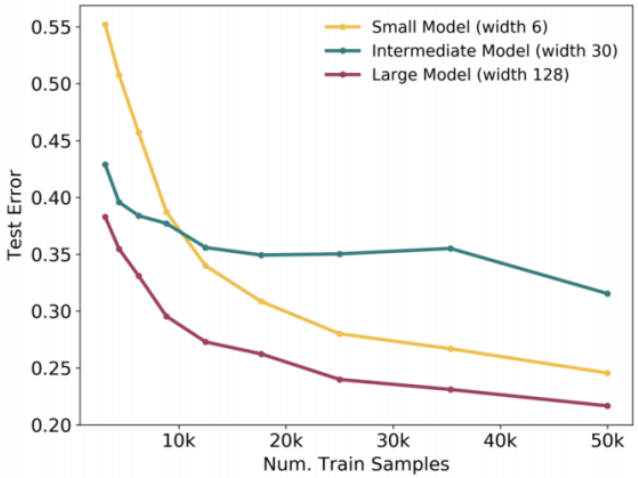
- В некоторых случаях объединение этих двух эффектов приводит к ухудшению качества модели

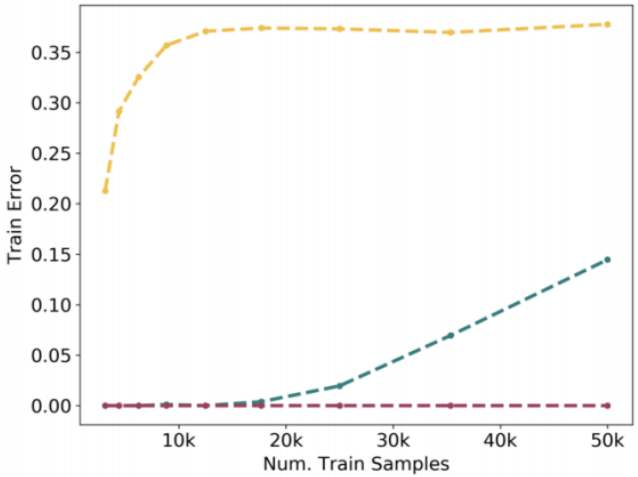


Sample-wise Double Descent

- Вне критической зоны увеличение выборки приводит к улучшению качества модели

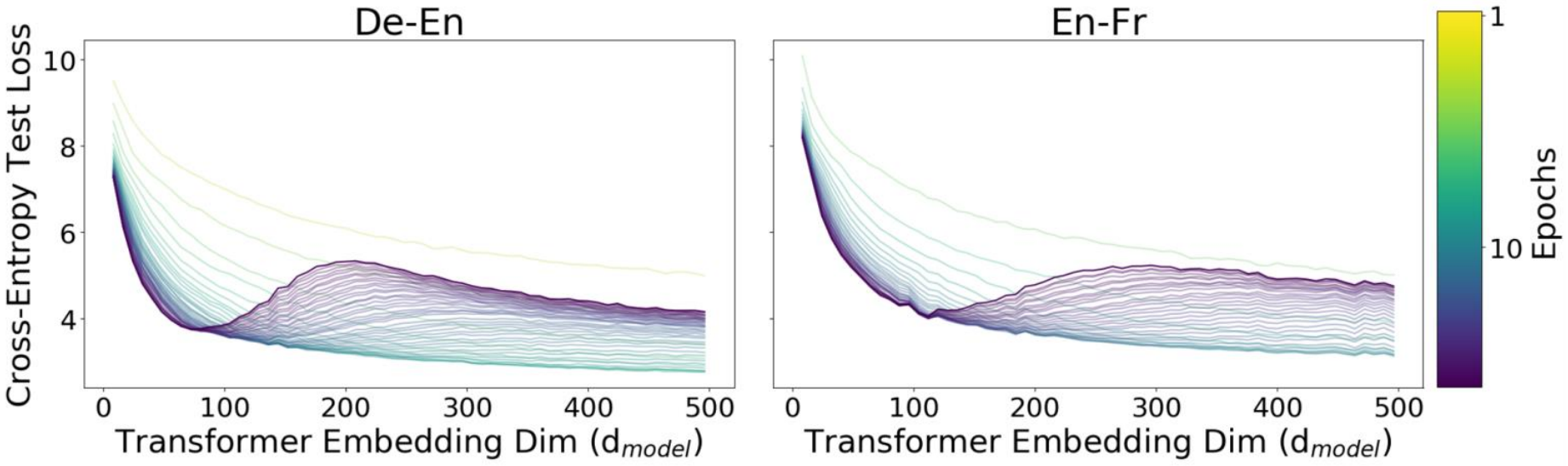






Early Stopping

- В основном при оптимальной ранней остановке не наблюдается Double Descent
- Это объясняется тем, что ранняя остановка предотвращает достижение нулевой тренировочной ошибки
- Увеличение размера выборки не ухудшает качество модели



Выводы

- Есть два режима обучения: классический и интерполяционный
- Классический режим имеет U-образый вид, интерполяционный - в целом убывающий
- Интерполяционный порог зависит от сложности модели, данных и процедуры обучения
- Если модель находится в диапозоне интерполяционного порога, то небольшие изменения модели и метода обучения могут привести к неожиданному поведению

Источники

- Reconciling modern machine-learning practice and the classical bias–variance trade-off
- Deep Double Descent: Where Bigger Models and More Data Hurt