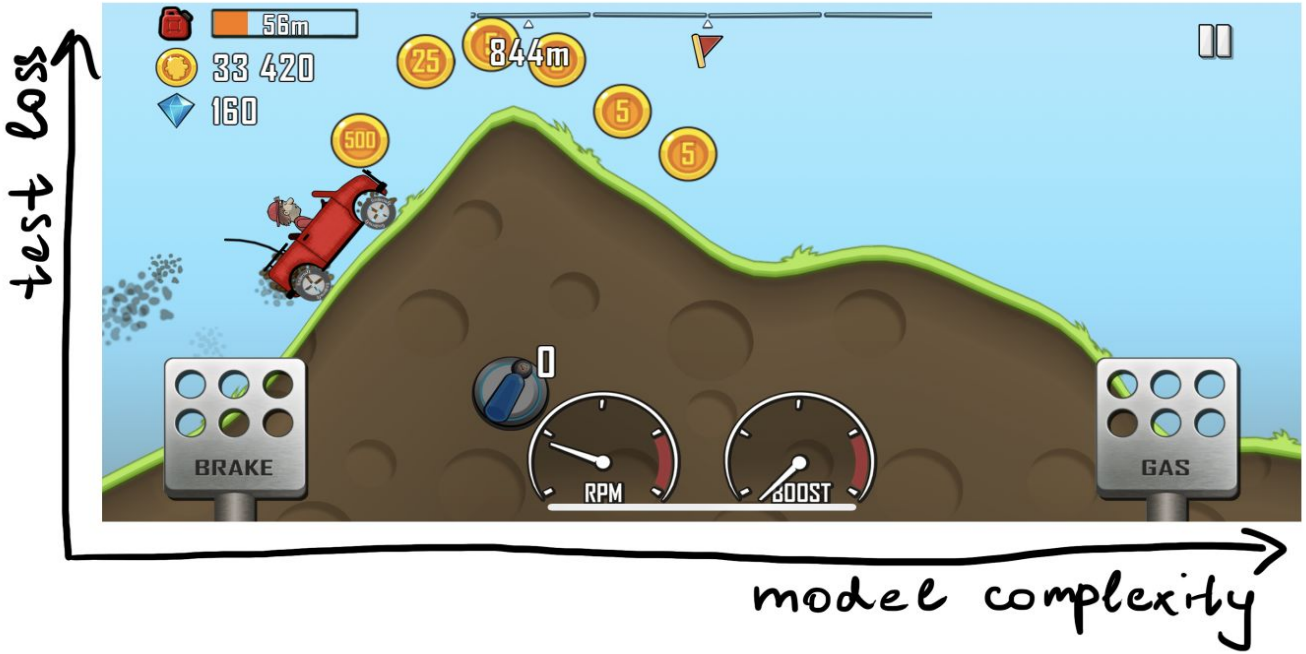
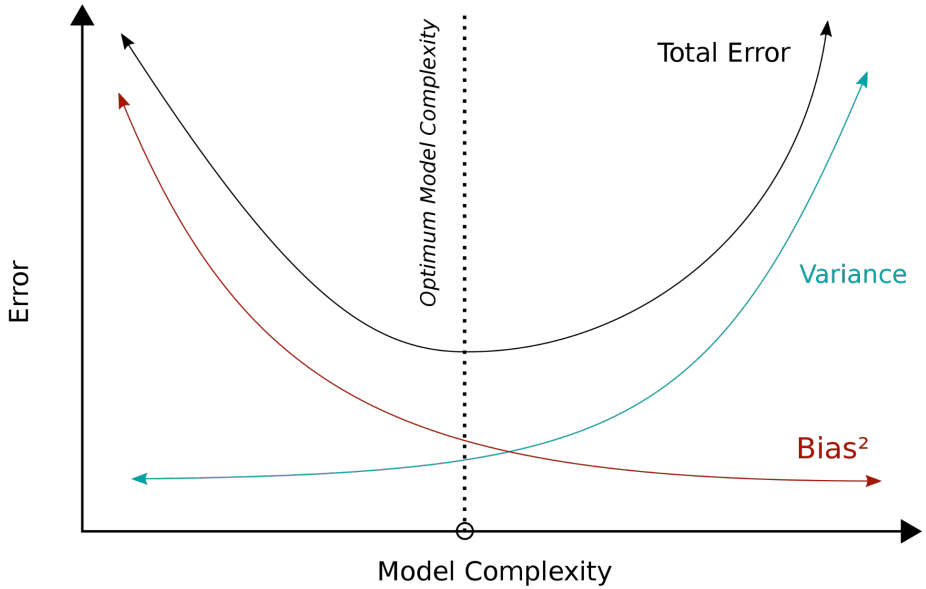
Double Descent



Данил
Шешеня
213

Bias-variance trade-off

- Теоретический концепт
- С ростом сложности модели падает смещение, но растет дисперсия



Житейская мудрость

- Не делать слишком сложные модели
- Тем не менее, большие модели существуют и хорошо работают
- Делать сложные модели!!
- Оба подхода верны

Effective Model Complexity (EMC)

Максимальное число элементов в обучающей выборке, при котором модель достигает ~0 ошибки

Пусть:

- $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ – выборка из распределения $\mathcal{D}$
- $\mathcal{T}(S)$ – процедура обучения
- $\varepsilon > 0$ (на практике ≈ 0.1)
- $\mathrm{Error}_S(M)$ – средняя ошибка модели M на обучающей выборке S

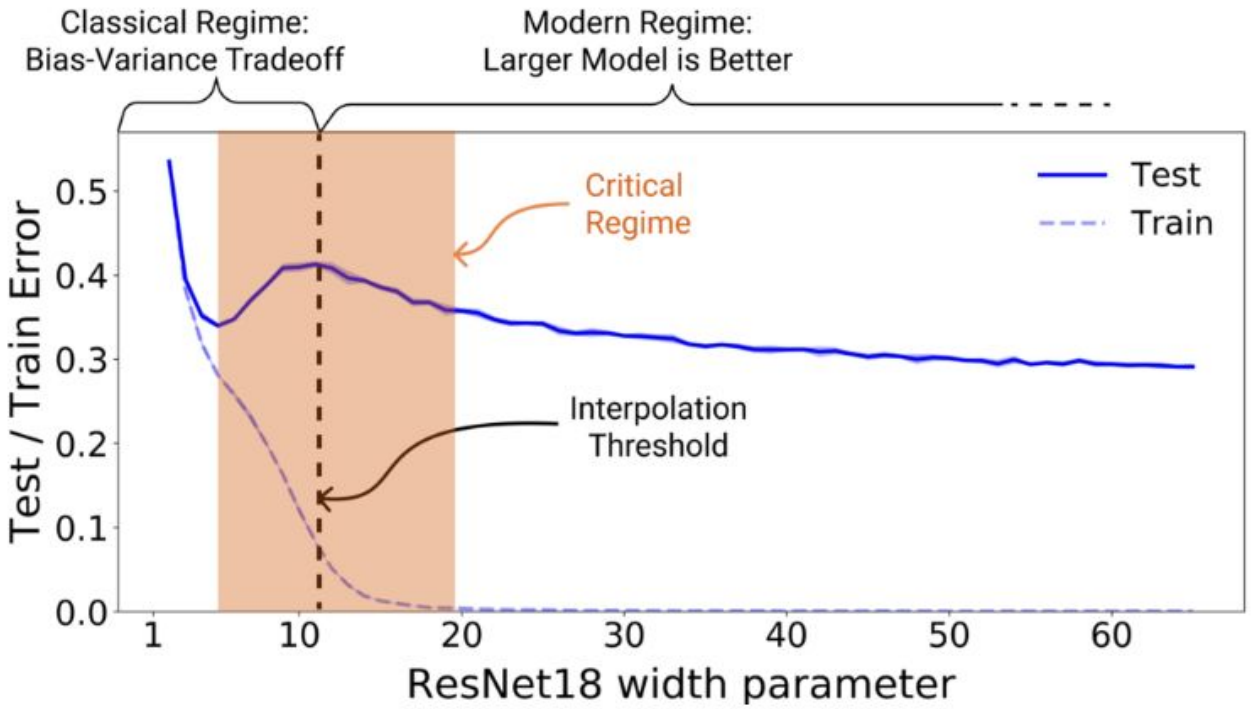
$$\mathrm{EMC}_{\mathcal{D},\varepsilon}(\mathcal{T}) := \max\{n \mid \mathbb{E}_{S\sim\mathcal{D}^n}[\mathrm{Error}_S(\mathcal{T}(S))] \leqslant \varepsilon\}$$

Зависит от распределения!

Гипотеза

Для любых $\mathscr{D}$, $\mathscr{T}$, $\varepsilon > 0$, если рассмотреть задачу классификации n элементов выборки

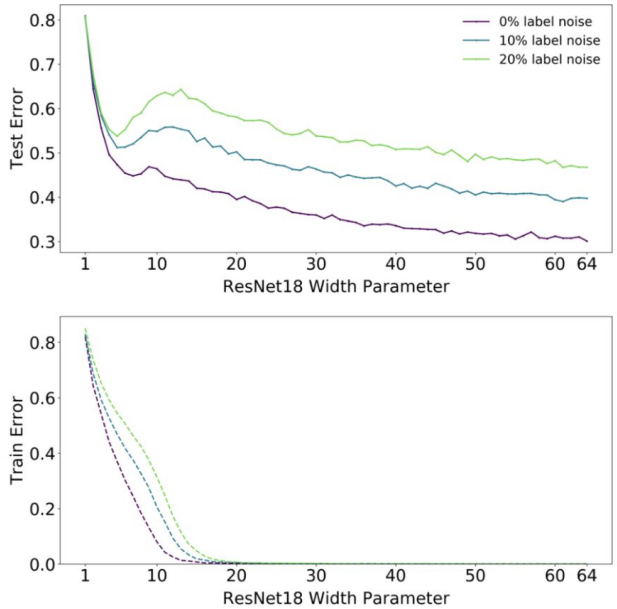
- *Недопараметризованный режим.* Если EMC << n, то с ростом EMC падает ошибка на тесте
- *Перепараметризованный режим.* Если EMC >> n, то с ростом EMC падает ошибка на тесте
- *Критически параметризованный режим.* Если EMC ~ n, то с ростом EMC ошибка на тесте падает или **растет**


Classical Regime:
Bias-Variance Tradeoff
Modern Regime:
Larger Model is Better
Critical
Regime
Test
Train
Interpolation
Threshold
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

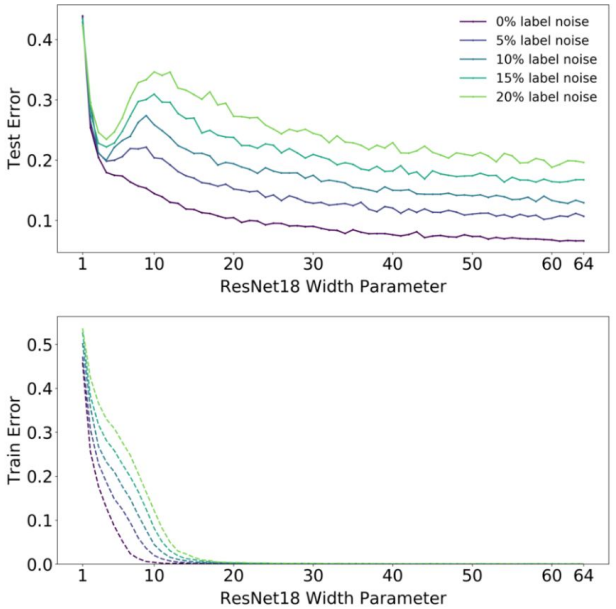
От чего зависит EMC?

- Архитектура модели (CNN, ResNet, Transformer)
- Параметры сети (ширина, число слоев и т.д)
- Регуляризация / аугментации / …
- Число эпох
- Оптимизатор (SGD, Adam)
- Уровень шума в данных

Model-Wise Double Descent

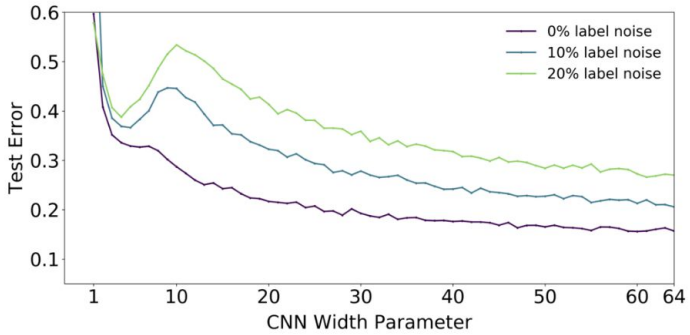


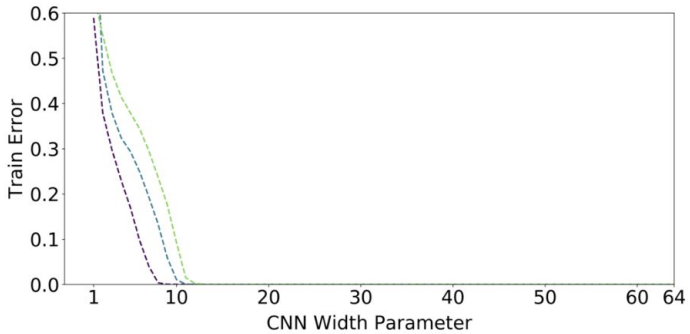
(a) **CIFAR-100.** There is a peak in test error even with no label noise.



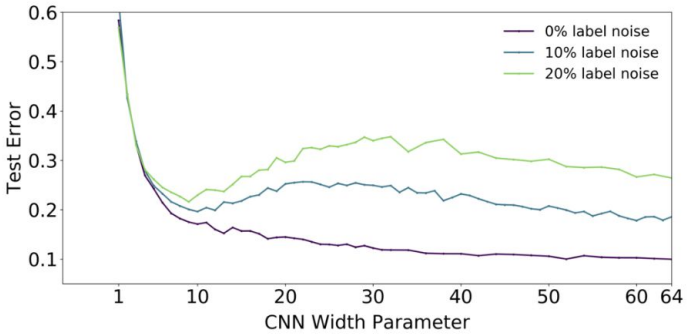
(b) **CIFAR-10.** There is a "plateau" in test error around the interpolation point with no label noise, which develops into a peak for added label noise.

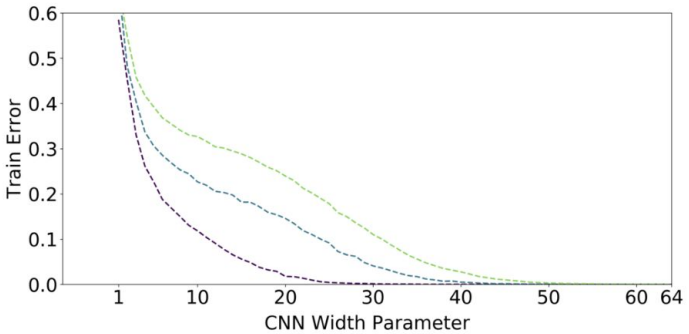
Model-Wise Double Descent





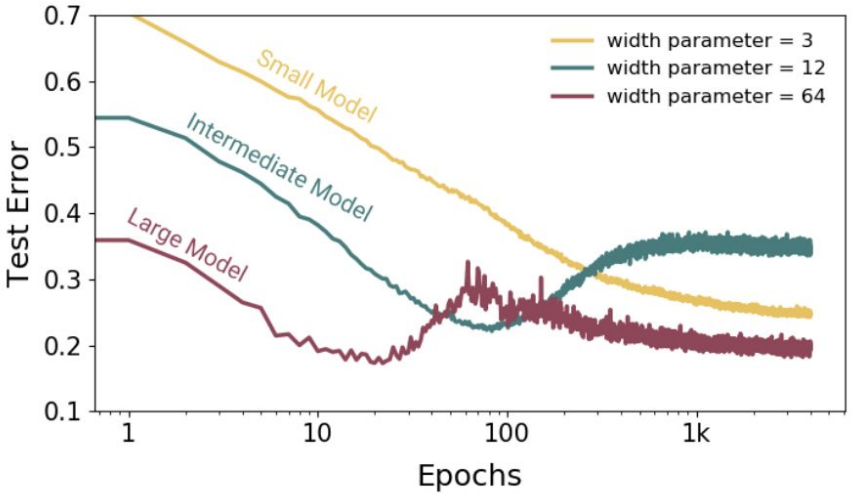
(a) Without data augmentation.

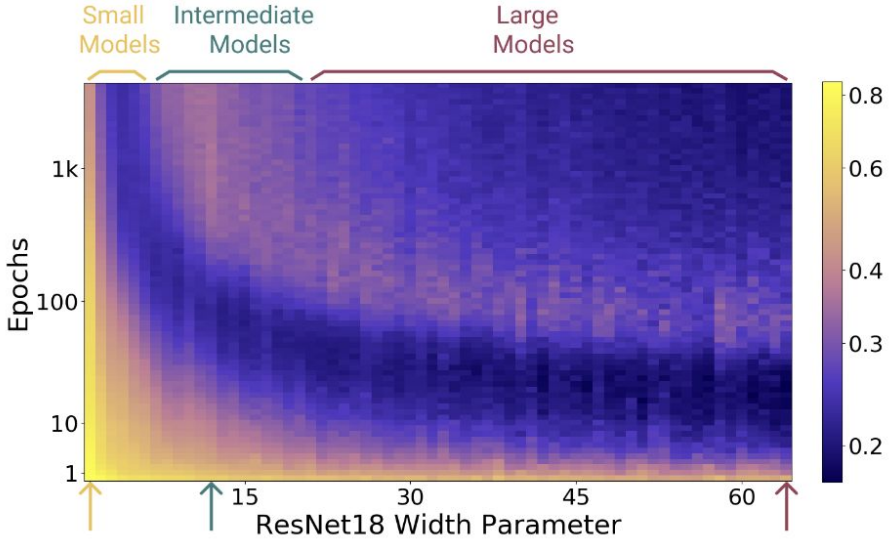




(b) With data augmentation.

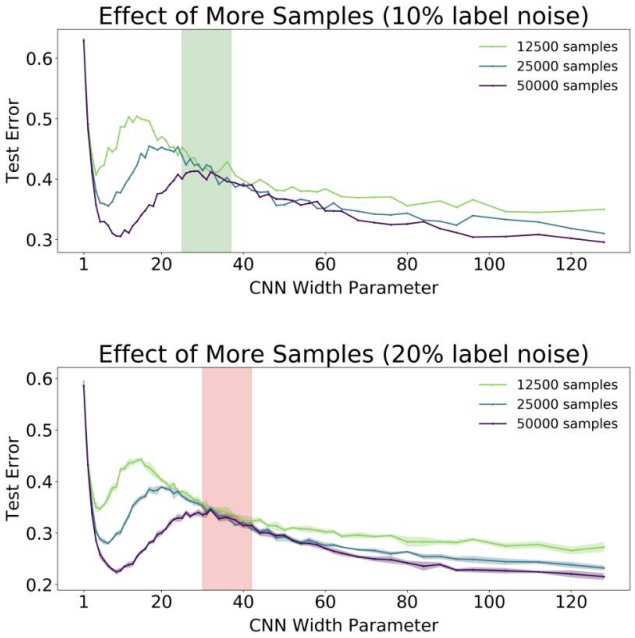
Epoch-Wise Double Descent

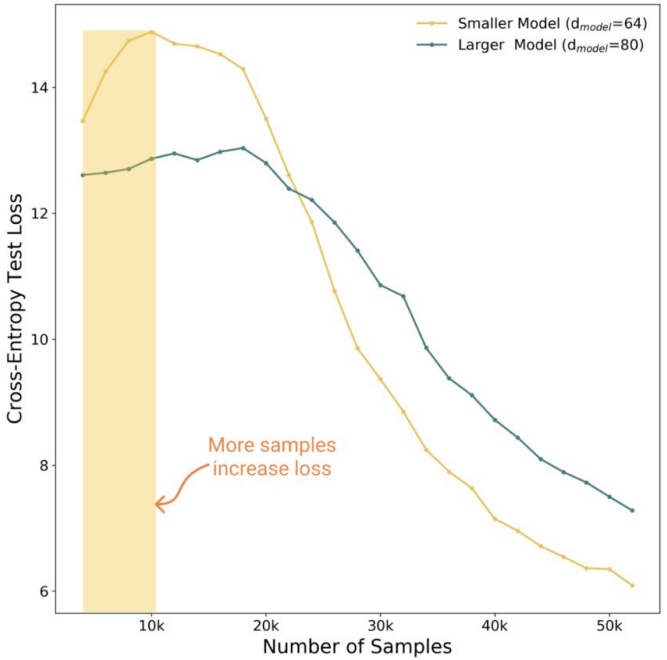




Sample-Wise Non-Monotonicity

Ситуации EMC ~ n можно достичь меняя n





Результаты

					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

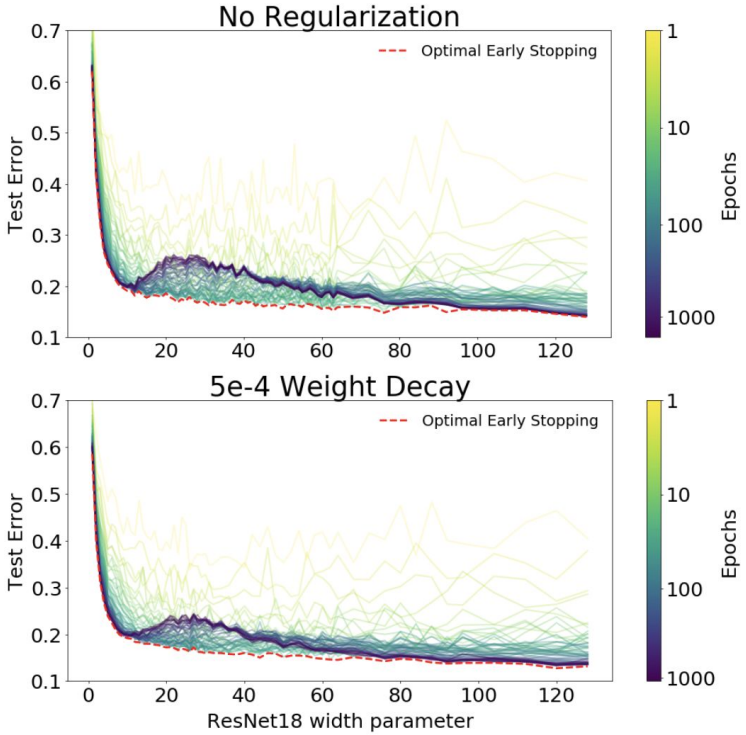
Попытка объяснения эффекта

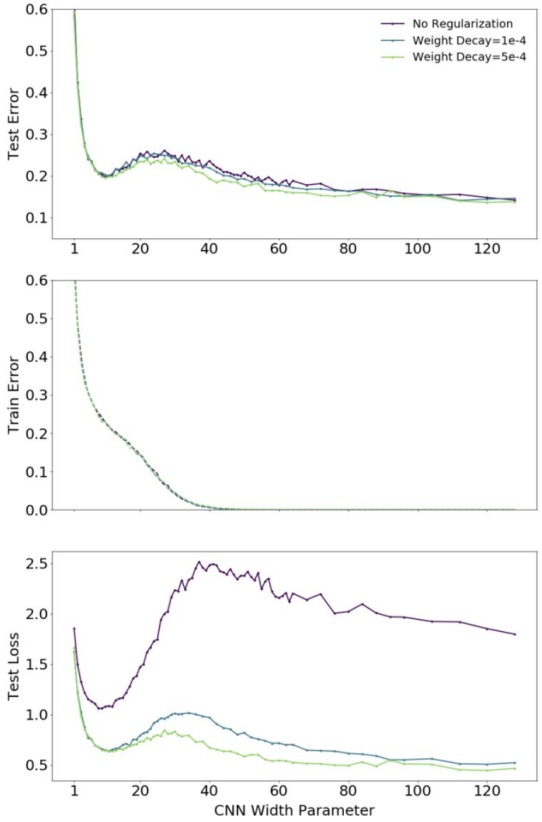
- По мере увеличения сложности модели, мы приближаемся к интерполяции обучающей выборки
- Такая интерполяционная модель единственна
- Она “шаткая”, и небольшой шум может ее сильно нарушить
- Но дальше есть много моделей, которые интерполируют обучающую выборку и сглаживают шумы

Заключение

- Эффект проявляется довольно устройчиво
- Точные причины эффекта никто не понимает
- Размер критической области тоже никто оценивать не умеет
- Большую роль играет зашумленность данных (model mis-specification)
- Если модель почти запомнила обучающую выборку, то небольшие изменения параметров могут привести к непредвиденным результатам
- Следить за ошибкой на обучающей выборке

Если успею


No Regularization
Optimal Early Stopping
Test Error
Epochs
1
10
100
1000
5e-4 Weight Decay
Optimal Early Stopping
Test Error
Epochs
ResNet18 width parameter


No Regularization
Weight Decay=1e-4
Weight Decay=5e-4
Test Error
Train Error
Test Loss
CNN Width Parameter

Еще картинки

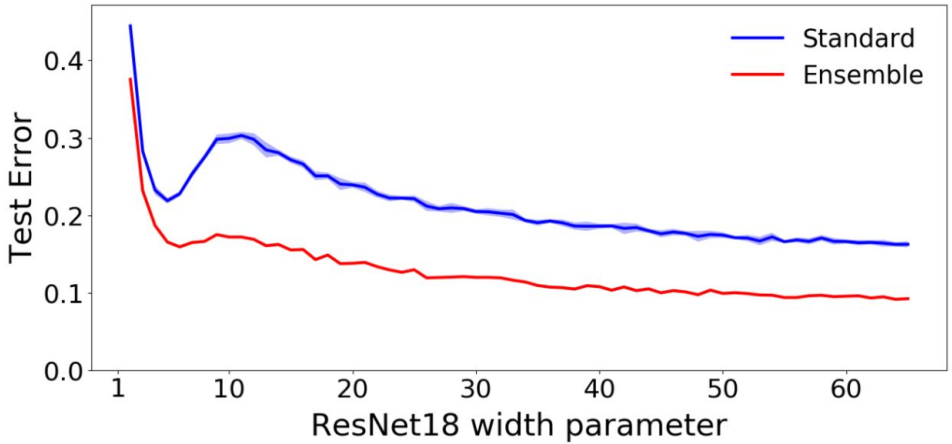
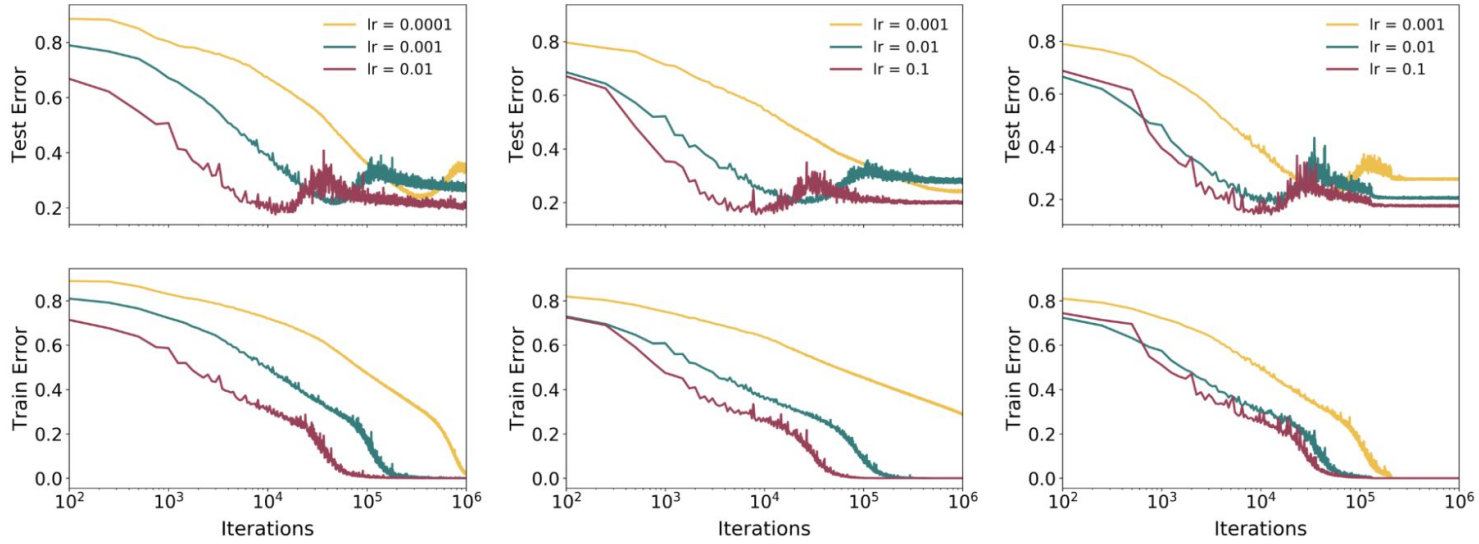


Figure 28: **Effect of Ensembling (ResNets, 15% label noise).** Test error of an ensemble of 5 models, compared to the base models. The ensembled classifier is determined by plurality vote over the 5 base models. Note that emsembling helps most around the critical regime. All models are ResNet18s trained on CIFAR-10 with 15% label noise, using Adam for 4K epochs (same setting as Figure 1). Test error is measured against the original (not noisy) test set, and each model in the ensemble is trained using a train set with independently-sampled 15% label noise.

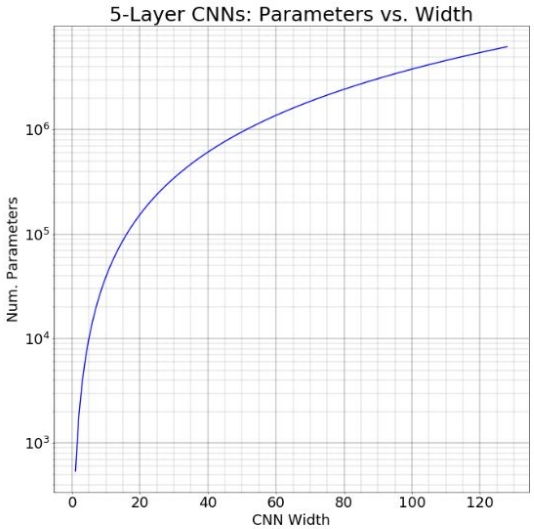
Сами разбирайтесь



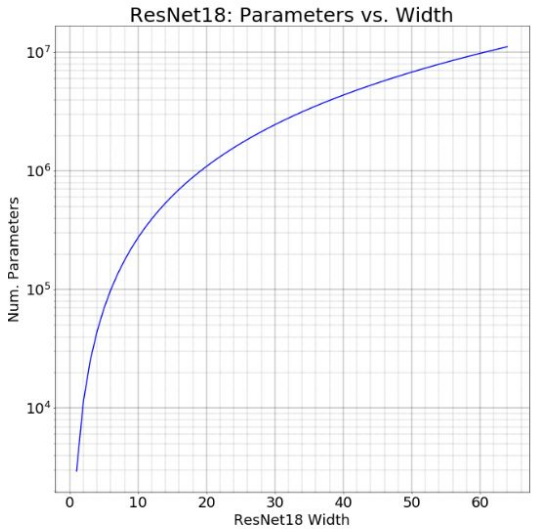
(a) Constant learning rate (b) Inverse-square root learning rate (c) Dynamic learning rate

Figure 17: **Epoch-wise double descent** for ResNet18 trained with SGD and multiple learning rate schedules

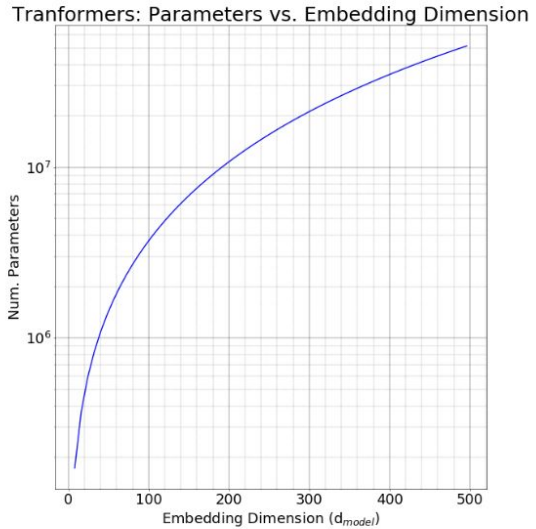
Ты думал я закончил?



(a) 5-layer CNNs



(b) ResNet18s



(c) Transformers