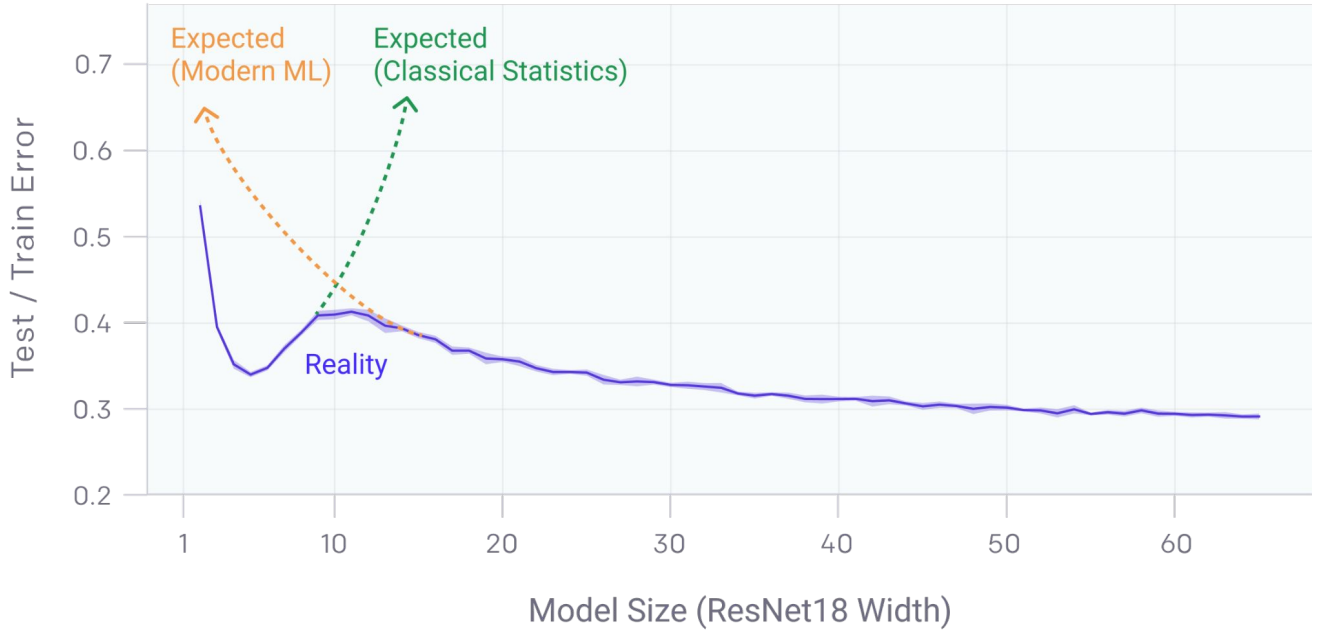
Deep Double Descent: Where Bigger Models and More Data Hurt

Терехова Юлия
Котельников Аким
Пак Ди Ун
Малафеев Михаил

Что это?



как работают современные нейронные сети

Effective model complexity

input a set $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$

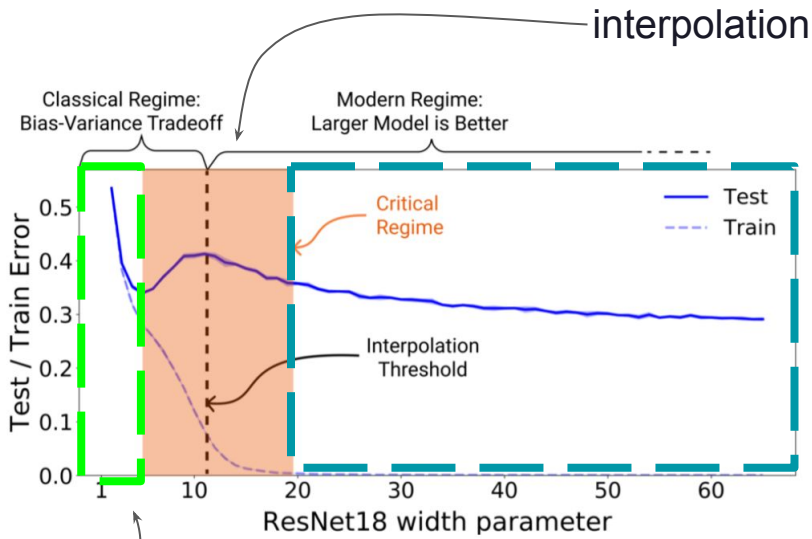
outputs a classifier $\mathcal{T}(S)$ mapping data to labels

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.*

Гипотеза



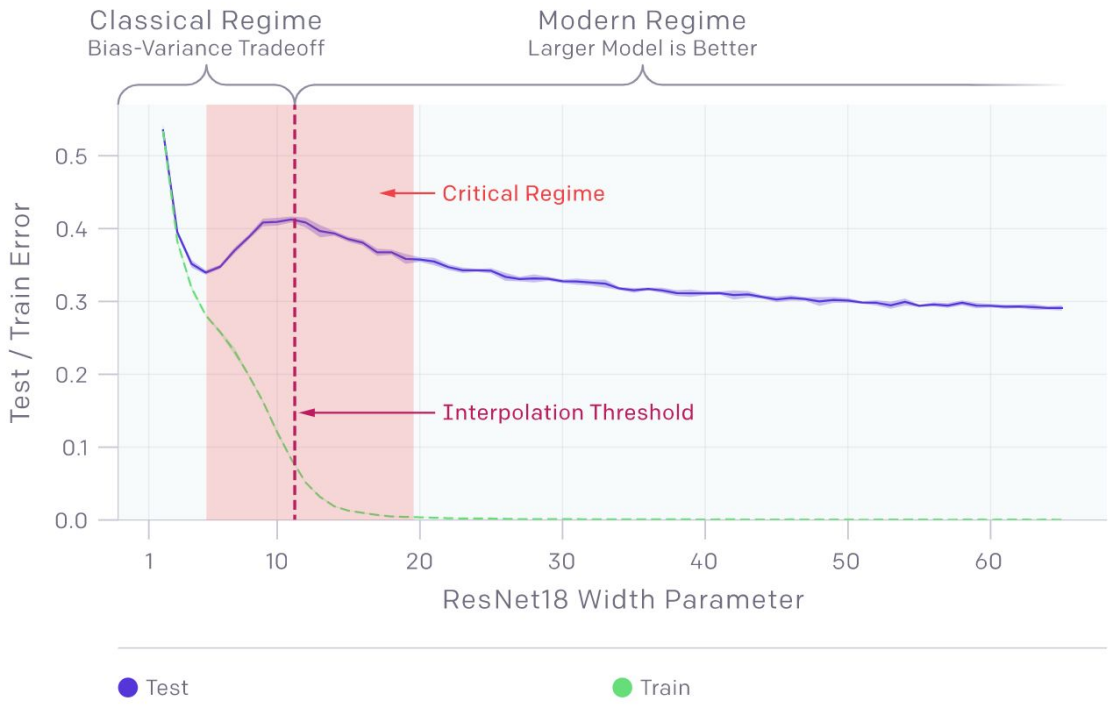
Under-paremeterized regime. *If* $\mathrm{EMC}_{\mathcal{D},\epsilon}(\mathcal{T})$ *is sufficiently smaller than* n*, any perturbation of* $\mathcal{T}$ *that increases its effective complexity will decrease the test error.*

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.*

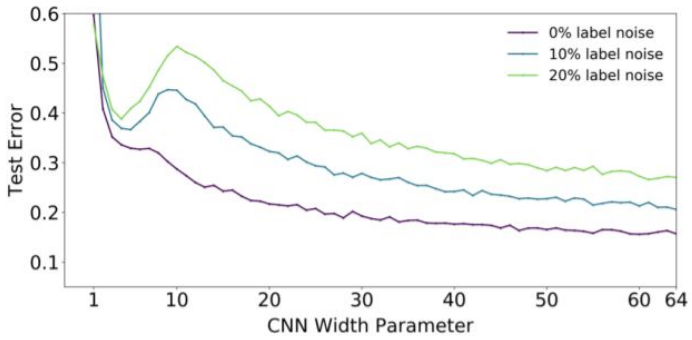
Critically parameterized regime. *If* $\mathrm{EMC}_{\mathcal{D},\epsilon}(\mathcal{T}) \approx n$*, then a perturbation of* $\mathcal{T}$ *that increases its effective complexity might decrease* **or increase** *the test error.*

Model-wise

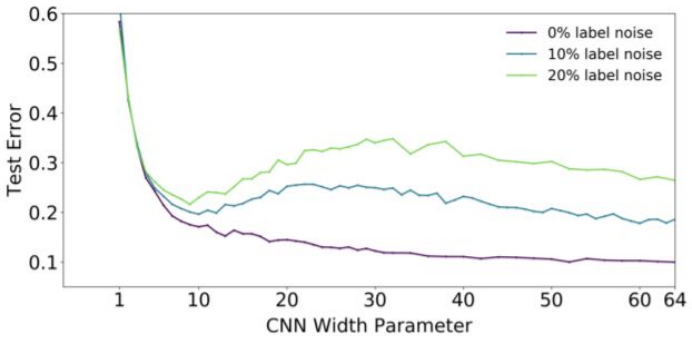
bigger models are worse



Эксперименты



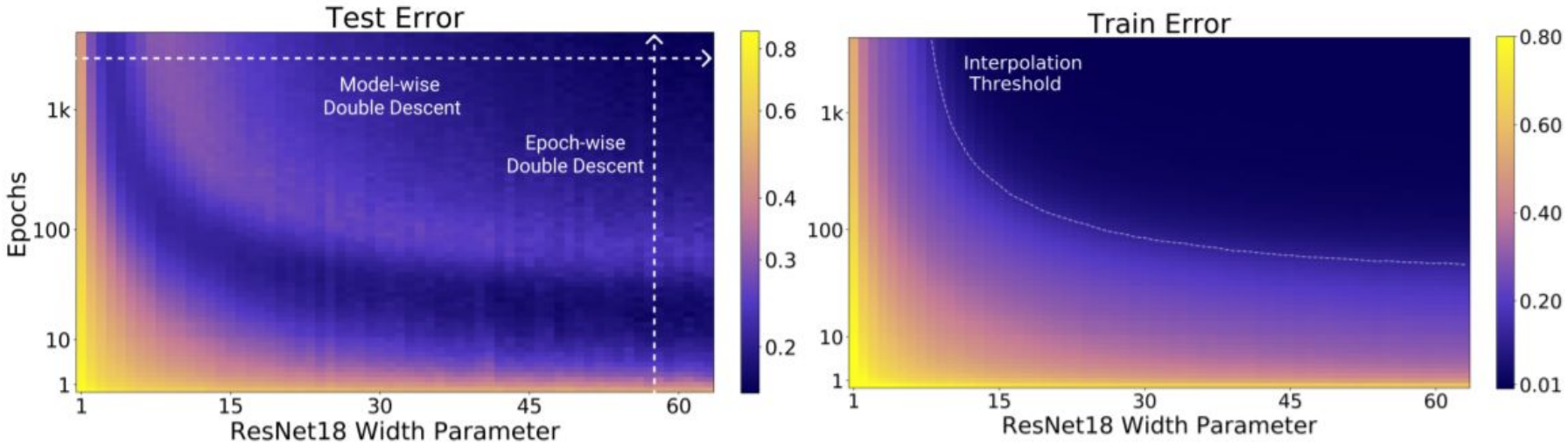
(a) Without data augmentation.



(b) With data augmentation.

Epoch-wise

training longer reverses overfitting



Эксперименты

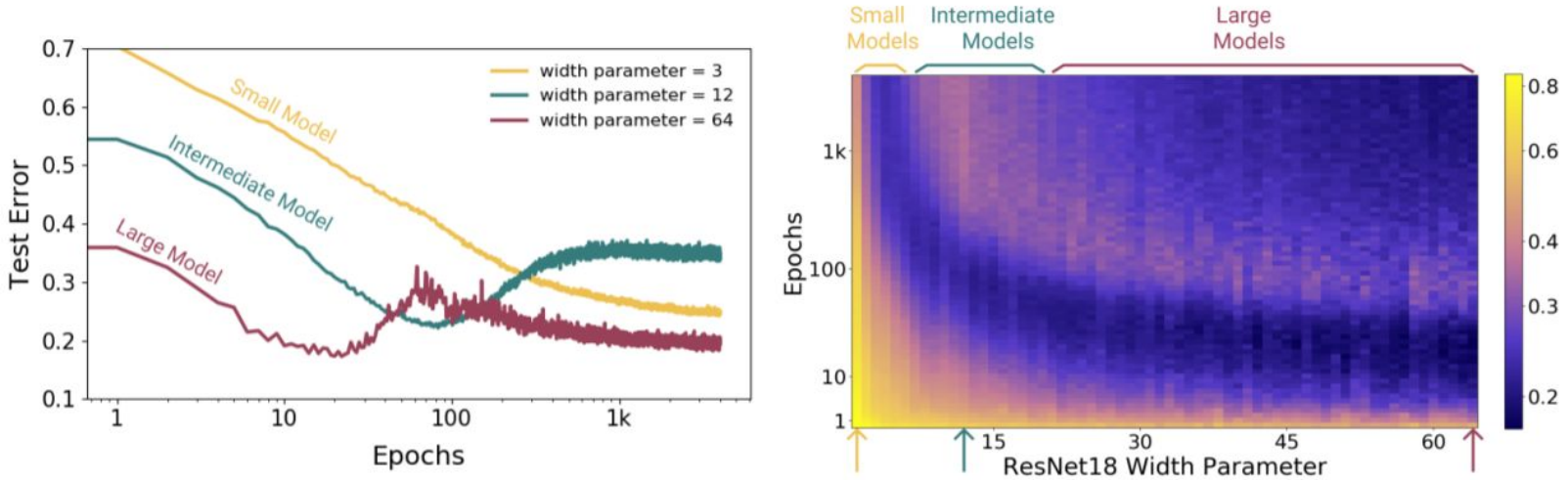
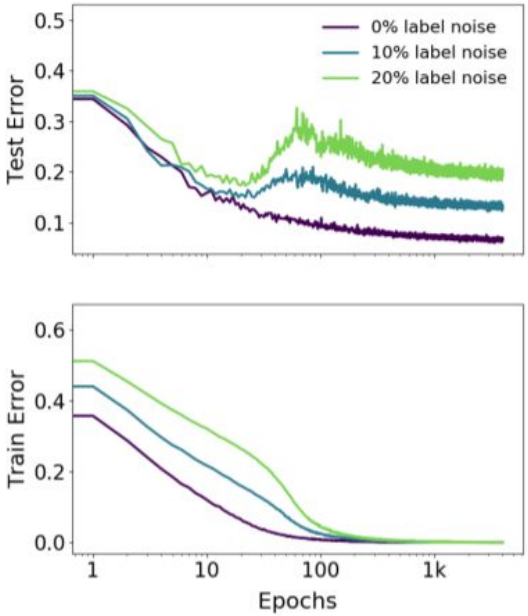
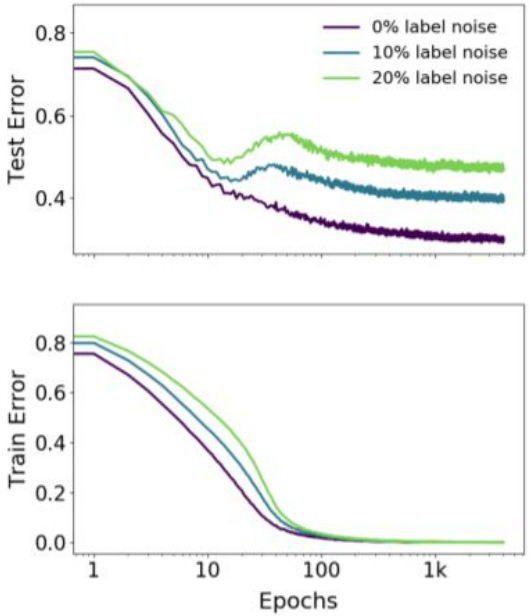


Figure 9: **Left:** Training dynamics for models in three regimes. Models are ResNet18s on CIFAR10 with 20% label noise, trained using Adam with learning rate 0.0001, and data augmentation. **Right:** Test error over (Model size $\times$ Epochs). Three slices of this plot are shown on the left.

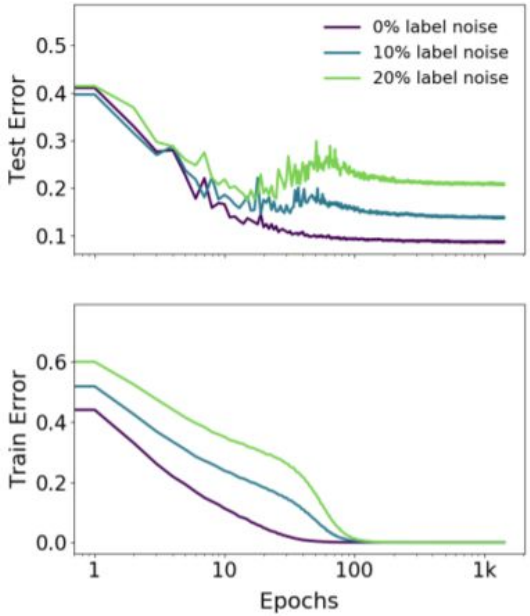
Эксперименты



(a) ResNet18 on CIFAR10.



(b) ResNet18 on CIFAR100.

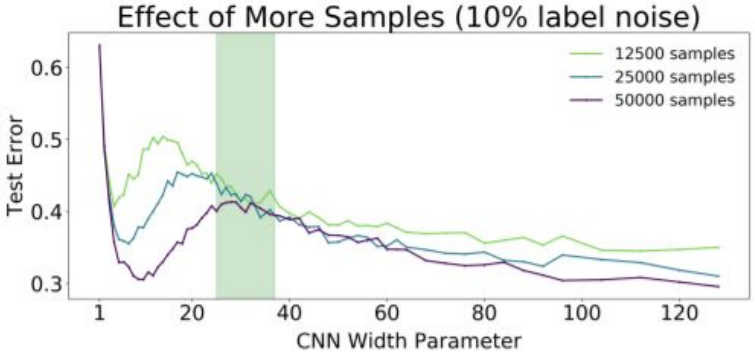


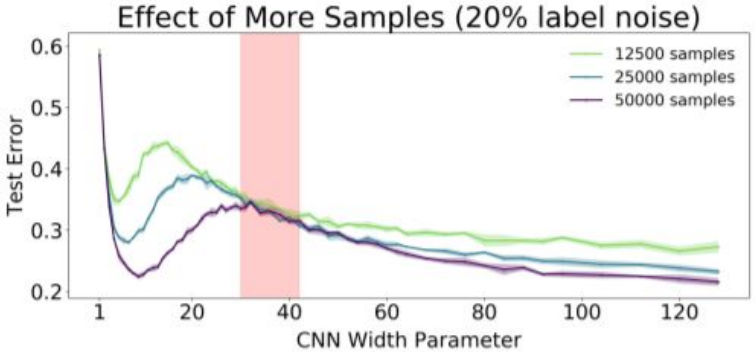
(c) 5-layer CNN on CIFAR 10.

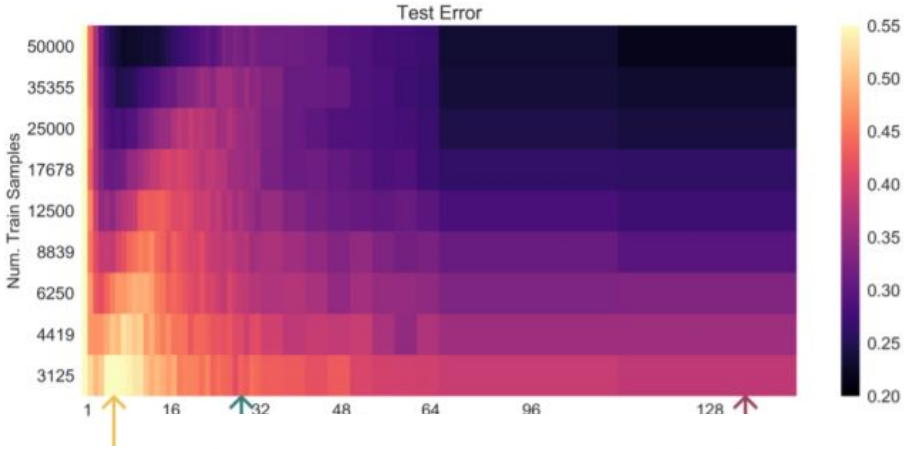
Sample-wise

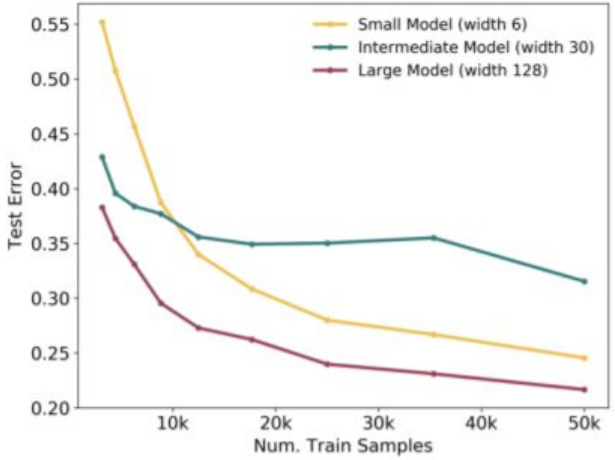
more samples hurts the performance
of the model

Эксперименты









Рецензия

Авторы статьи проделали эксперименты, изучая феномен “Deep Double Descent”, при котором у моделей большого размера ошибка начинает расти в какой-то момент, но потом снова падает. Этот эффект зависит как от размера модели, так и от количества эпох. Эксперименты показывают, что классический bias-variance trade-off не выполняется для больших моделей. Также были изучены интересные эффекты, связанные с тем, что увеличение количества данных может увеличить ошибку на тесте.

Авторы ввели новую понятие Effective Model Complexity обучающего процесса, которое обозначает максимальное число сэмплов, на которых обучающий процесс достигает нулевой ошибки на этой выборке. Также авторы выдвинули гипотезу основываясь на этом понятии, которая говорит о взаимодействии между EMC и ошибки на тесте

Рецензия

Плюсы:

- Интересные результаты с epoch-wise ddd и sample-wise non-monotonicity
- Много подробных экспериментов, графиков, которые хорошо объяснены
- Статья понятно написана, легко читается

Минусы:

- Эффект model-wise ddd был исследован ранее в нескольких статьях (пример). Тем не менее в данной статье эксперименты гораздо обширнее и для более новых архитектур

Воспроизводимость:

- Смысл экспериментов понятен, их несложно проделать, но их много

Оценка

- Оценка: 8/10
- Уверенность 4/5

Исследование

- Публикации
- Авторы
- Опорные статьи
- Конкуренты
- Продолжение работы

Публикации

1. ICLR2020 Conference
2. До подачи на конференцию данная работа была опубликована в блоге OpenAI с последующими комментариями от авторов в другом исследовательском блоге

Авторы

Preetum Nakkiran

- Bachelor in Electrical Engineering and Computer Science (University of California, Berkeley)
- PhD in Computer Science (Harvard University)
- Postdoc 2021-now (University of California, San Diego)
- Ревьювер NeurIPS, ICLR, JMLR, Distill, STOC, CRYPTO, ITCS, IEEE Transactions on Information Theory
- H-index: 12
- 891 citations

Авторы

Gal Kaplun

- Bachelor Hebrew in Mathematics and Computer Science (University of Jerusalem)
- Self-driving car firm Mobileye, 2017-2018
- PhD in Computer Science, 2020-now (Harvard University)
- H-index: 4
- 360 citations

Авторы

Yamini Bansal

- Bachelor and masters in IIT Bombay
- PhD in Computer Science (Harvard University)
- Paper “Revisiting Model Stitching to Compare Neural Representations”, NeurIPS 2021
- H-index: 4
- 565 citations

Авторы

Tristan Yang

- Sophomore in Computer Science (Harvard University)

Авторы

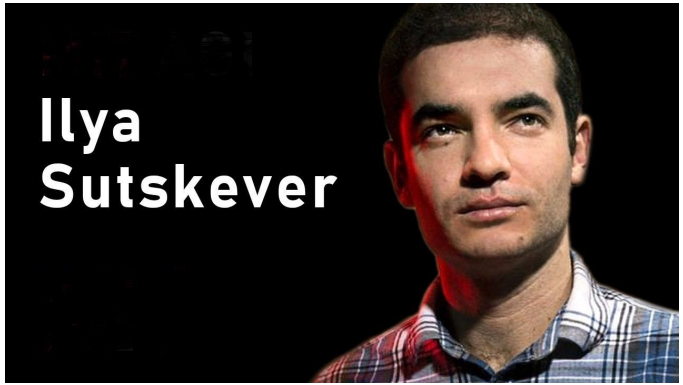
Boaz Barak

- PhD Weizmann Institute of Science
- Postdoctoral at the Institute for Advanced Study (Princeton)
- Ex associate professor at Princeton University
- Ex principal researcher at Microsoft Research New England
- Professor in Computer Science (Harvard University)
- H-index: 47
- 13660 citations

Авторы

Ilya Sutskever

- Bachelor, master and PhD (University of Toronto)
- Chief scientist of OpenAI
- Co-inventor of AlexNet
- One of the authors of the AlphaGo
- 35 Innovators Under 35 from MIT ranking in 2015
- H-index: 66
- 277531 citations



На что опирается

1. *Chiyuan Zhang et al*, “Understanding deep learning requires rethinking generalization.”. Overparametrised DNN achieve decent generalization
2. *Yanping Huang et al*, “GPipe: Efficient Training of Giant Neural Networks using Pipeline Parallelism”. Scaling any neural network with parallelism.
3. *Christian Szegedy et al*, “Going Deeper with Convolutions”. Very deep NN Inception.

Конкуренты

1. *Madhu S. Advani, Andrew M. Saxe*, “High-dimensional dynamics of generalization error in neural networks”. Overparametrised DNN double descent over time
2. *Stefano Spigler et al*, “A jamming transition from under- to over-parametrization affects loss landscape and generalization”. Loss landscape of over-under-parametriszed models
3. *Mikhail Belkin et al*, “Reconciling modern machine learning practice and the bias-variance trade-off”. Introduction into double descent

Продолжение работы

1. *Preetum Nakkiran et al*, “Optimal Regularization Can Mitigate Double Descent”. Eliminate DDD effect from training with proper regularization.
2. *Stéphane d'Ascoli et al*, “Double Trouble in Double Descent : Bias and Variance(s) in the Lazy Regime”. Bias-variance decomposition theory in Double Descent
3. *Stéphane d'Ascoli et al*, “Triple descent and the two kinds of overfitting: Where & why do they appear?”. Double peak in loss when N is equal to D and P at the same time in noisy task
4. *Ekaterina Lobacheva , Nadezhda Chirkova , Maxim Kodryan , Dmitry Vetrov*, “On Power Laws in Deep Ensembles”. ICML2021