Understanding Deep Learning Requires Rethinking Generalization

- ICLR 2017, best papers award
- Authors: Zhang, Bengio, Hardt, Racht, Vinyals
- Presenter: Hossein Azizpour

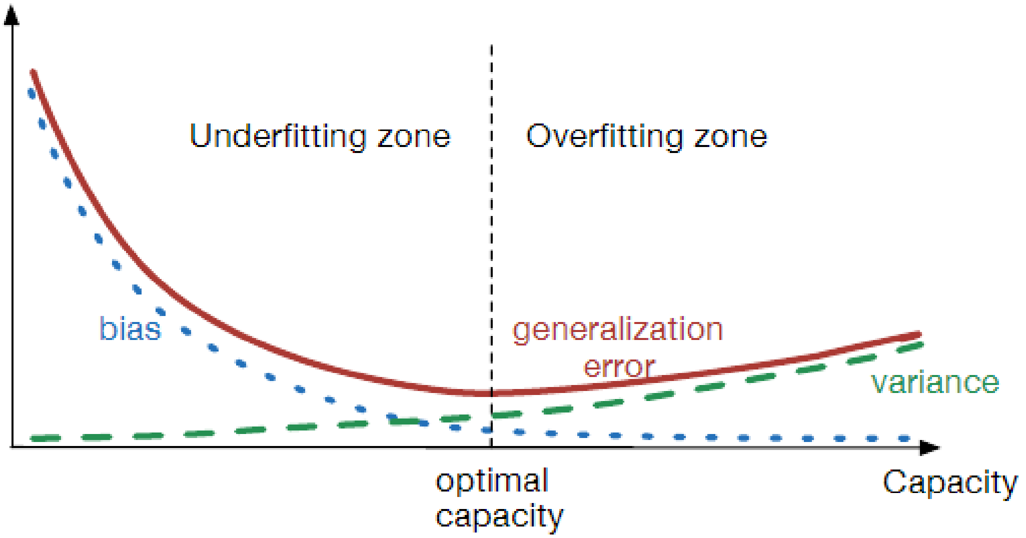
In a nutshell

Modern deep networks usually achieve low generalization error

Why?!

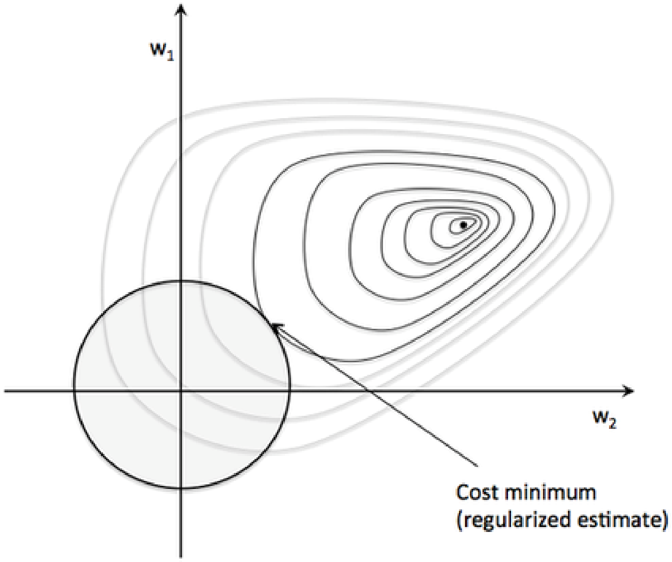
In a nutshell

- Statistical learning theory on low generalization error:
  - Properties of the model family and training procedure
    - Upper bound on generalization error



In a nutshell

- Statistical learning theory on low generalization error:
  - Properties of the model family and training procedure
    - Upper bound on generalization error
  - Regularization



Cost minimum
(regularized estimate)

In a nutshell

Study to see whether *regularization* or theories on the *generalization bound* explain the performance of deep networks

In a nutshell

- Experiments 1: (many) deep architectures can fit the same dataset they learn with low generalization error, even with random labels

Statistical learning theories cannot explain the generalization abilities of deep networks

In a nutshell

- Experiments 2: Removing all regularization practices of a modern deep network does not fundamentally cripple the generalization.

Rregularization is not a necessary factor for low generalization error in deep networks

In a nutshell

Thus,

we need to rethink “generalization” when dealing with deep networks!

Experiments

Experiments

- Datasets:

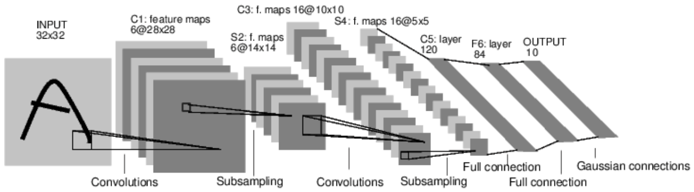
- Cifar10: 60K images, 32x32, 10 classes (animals and vehicles)

- ImageNet: 1.3M images, 299x299, 1000 classes

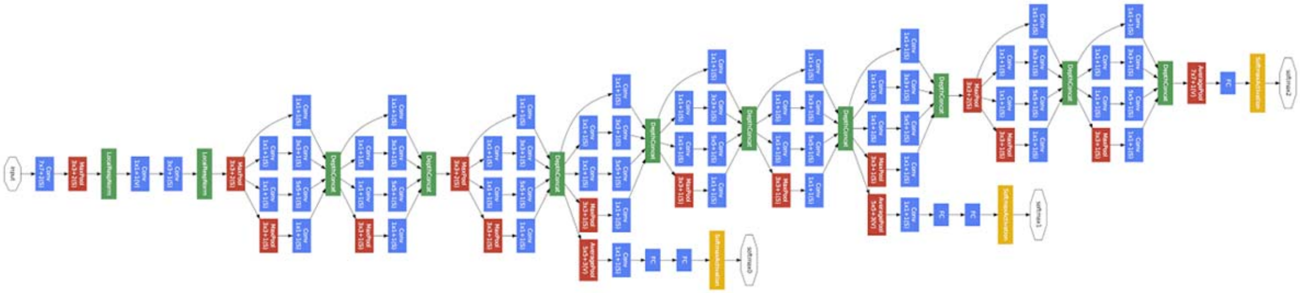
Experiments

- Architectures

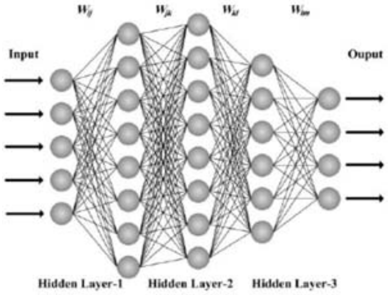
Modified AlexNet for Cifar10 (smaller)



Inception V4



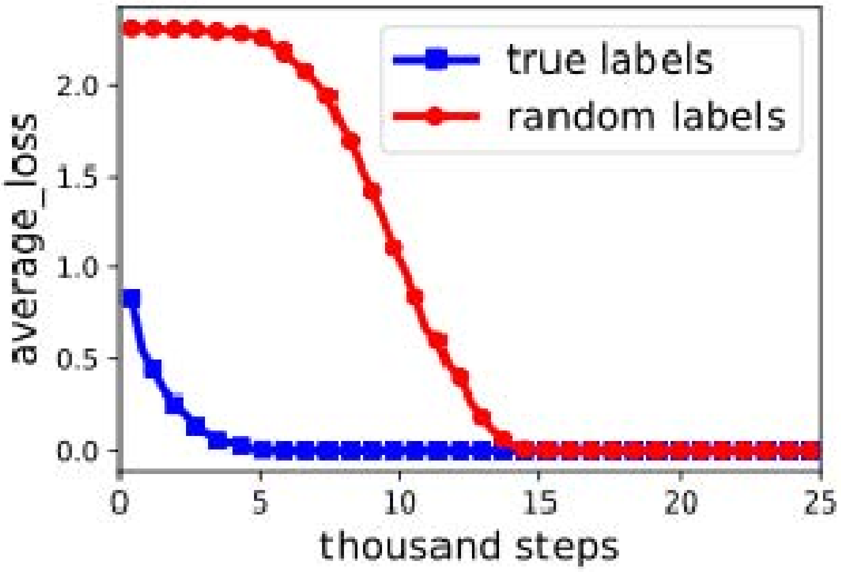
MLP

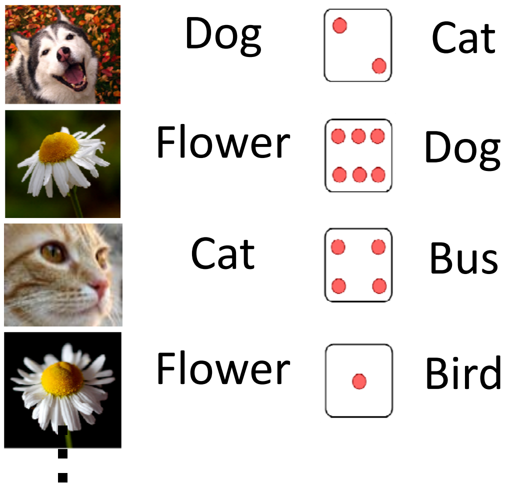


Experiments (1)

Randomization Test

Assign random labels to CIFAR10 images



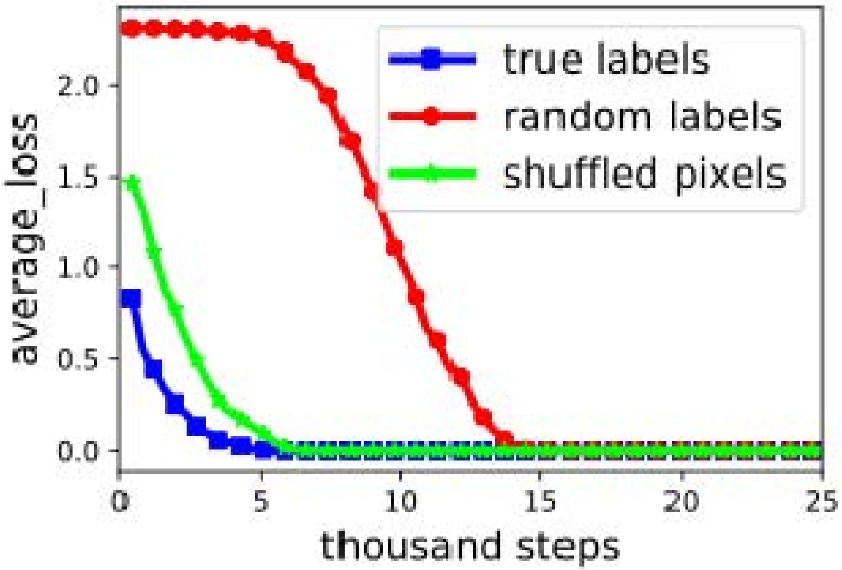


	Dog		Cat
	Flower		Dog
	Cat		Bus
	Flower		Bird

Experiments (1)

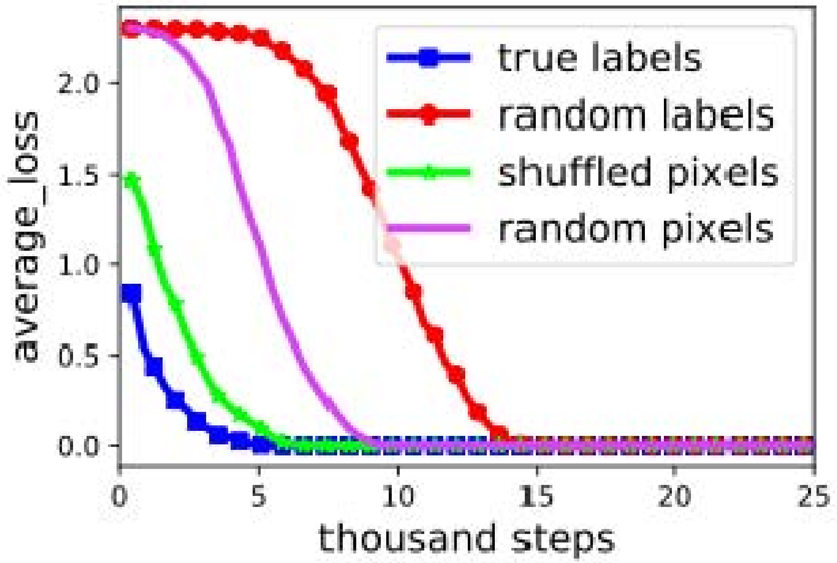
Randomization Test

Same random permutation applied to pixels of CIFAR10 images



Experiments (1)
Randomization Test

Different random permutation applied to pixels of CIFAR10 images

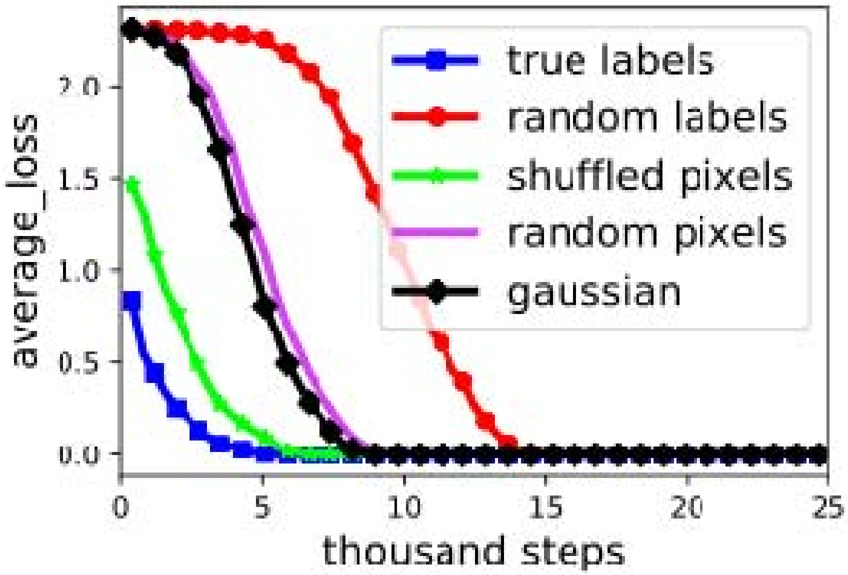


Experiments (1)

Randomization Test

Totally Gaussian random pixels with

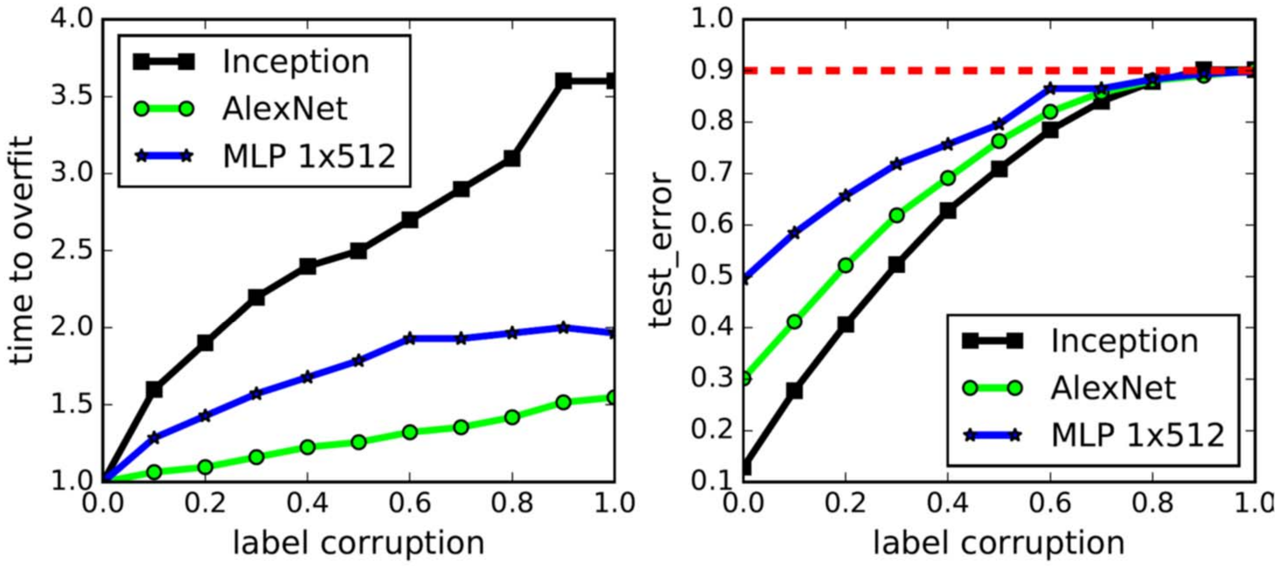
mean/variance from CIFAR10 images



Experiments (1)

Randomization Test

Soft Label Corruption



(b) convergence slowdown

(c) generalization error growth

“neural networks are able to capture the remaining signal in the data, while at the same time fit the noisy part using brute-force.”

Experiments (1)

some ***claims***

Refuted

"The effective capacity of neural networks is sufficient for *memorizing* the entire data set"

Experiments (1)

some ***claims***

“Even optimization on random labels remains easy. In fact, *training time increases only by a small constant* factor compared with training on the true labels.”

Not always

Experiments (1)

some ***claims***

“Randomizing labels is solely a data transformation, leaving all other properties of the *learning problem unchanged*”

vague

Experiments (1)

some ***claims***

The network learns different labeling of same data perfectly → shatters the data disregarding its labeling; Thus,

Rademacher complexity and VC dimension only offer unusable upper bound on the generalization error.

Argued

Experiments (2)

Turn off regularization

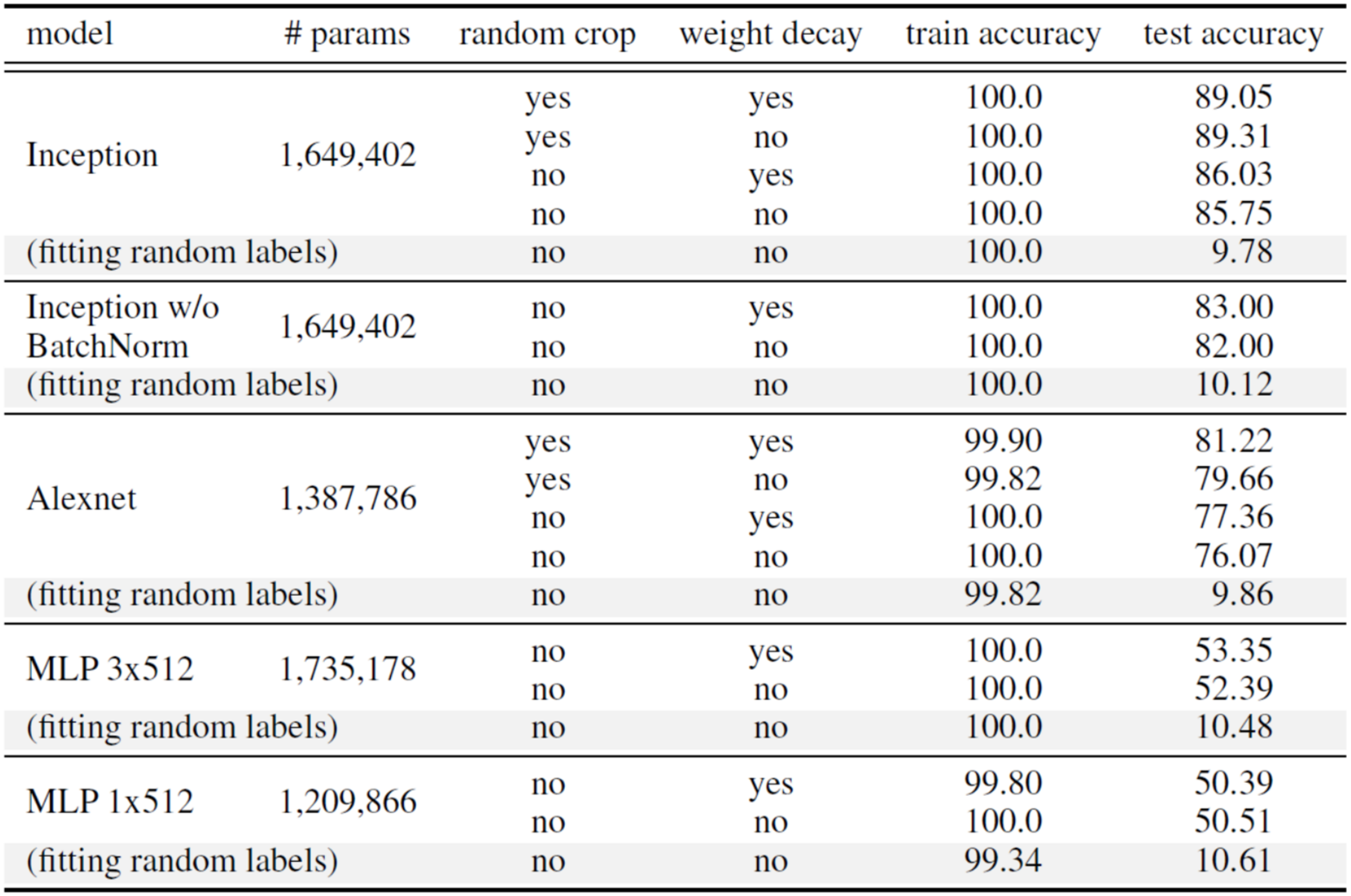
A REGULARIZER is ~~a mechanism~~ anything that ~~constrain the model or empower the data.~~ hurt the training Process.

Implicit vs. explicit regularization

If they are specifically designed for regularization

Experiments (2)

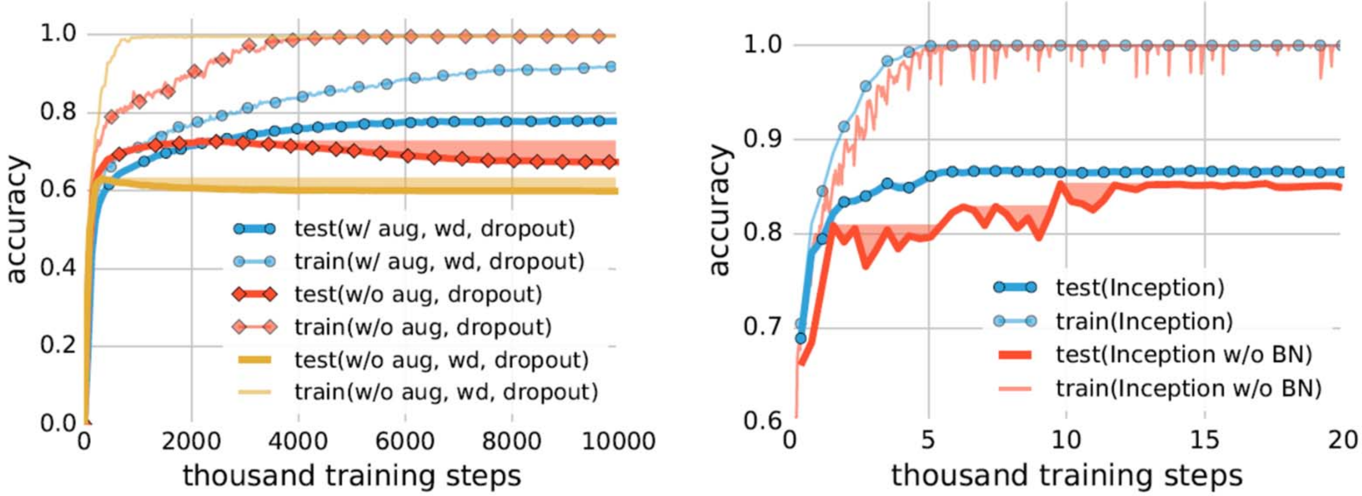
Turn off regularization



model	# params	random crop	weight decay	train accuracy	test accuracy
Inception	1,649,402	yes	yes	100.0	89.05
		yes	no	100.0	89.31
		no	yes	100.0	86.03
		no	no	100.0	85.75
(fitting random labels)		no	no	100.0	9.78
Inception w/o BatchNorm	1,649,402	no	yes	100.0	83.00
		no	no	100.0	82.00
(fitting random labels)		no	no	100.0	10.12
Alexnet	1,387,786	yes	yes	99.90	81.22
		yes	no	99.82	79.66
		no	yes	100.0	77.36
		no	no	100.0	76.07
(fitting random labels)		no	no	99.82	9.86
MLP 3x512	1,735,178	no	yes	100.0	53.35
		no	no	100.0	52.39
(fitting random labels)		no	no	100.0	10.48
MLP 1x512	1,209,866	no	yes	99.80	50.39
		no	no	100.0	50.51
(fitting random labels)		no	no	99.34	10.61

Experiments (2)

Turn off regularization



(a) Inception on ImageNet

(b) Inception on CIFAR10

Experiments (2)

claim

Explicit regularization may improve generalization performance, but is neither necessary nor by itself sufficient for controlling generalization error.

Some Theorems

Theorem (1)

expressivity of deep nets

“a very simple two-layer ReLU network with p = 2n+d parameters can express any labeling of any sample of size n in d dimensions.”

Theorem (1)

SGD as regularization

“For *linear models*, ***initialized at zero***, SGD always converges to a solution with small norm.”

ICLR 2017: ENTROPY-SGD: BIASING GRADIENT DESCENT INTO WIDE VALLEYS
NIPS 1999: SIMPLIFYING NEURAL NETS BY DISCOVERING FLAT MINIMA
ICLR 2017: On large-batch training for deep learning: Generalization gap and sharp minima.

Other works

DEEP NETS DON’T LEARN VIA MEMORIZATION

David Krueger, …, Aaron Courville

ICLR 2017 workshops

“Deep neural networks (DNNs) do not achieve their performance by memorizing training data, they learn a simple available hypothesis that fits the finite data samples.“

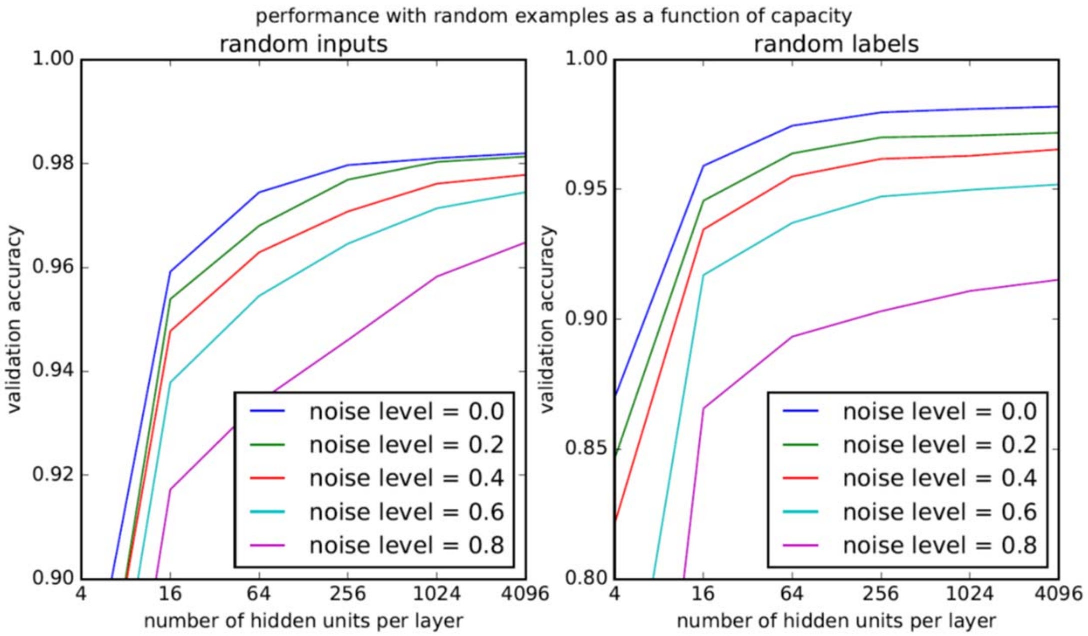
Other works

DEEP NETS DON'T LEARN VIA MEMORIZATION

David Krueger, …, Aaron Courville

ICLR 2017 workshops

"more capacity is needed to fit noise"



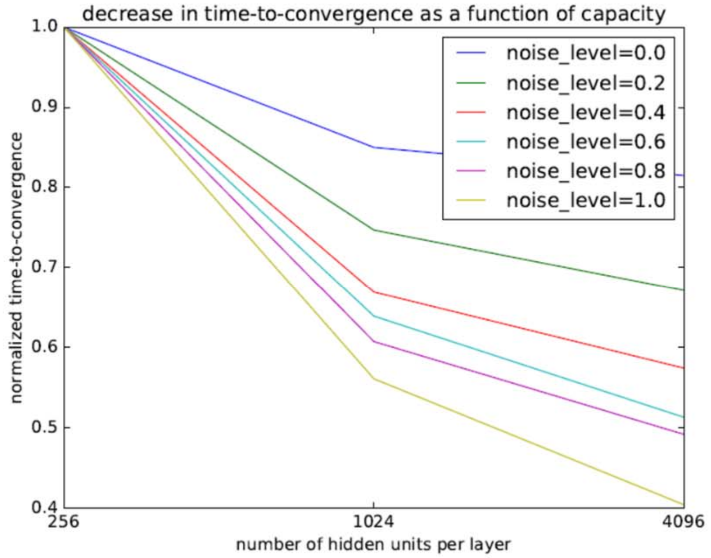
Other works

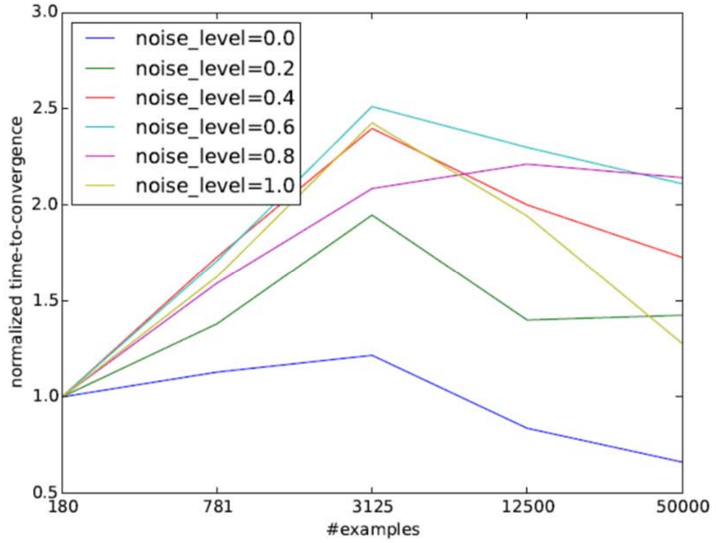
DEEP NETS DON’T LEARN VIA MEMORIZATION

David Krueger, …, Aaron Courville

ICLR 2017 workshops

“time to convergence is longer for random labels, but shorter for random inputs”





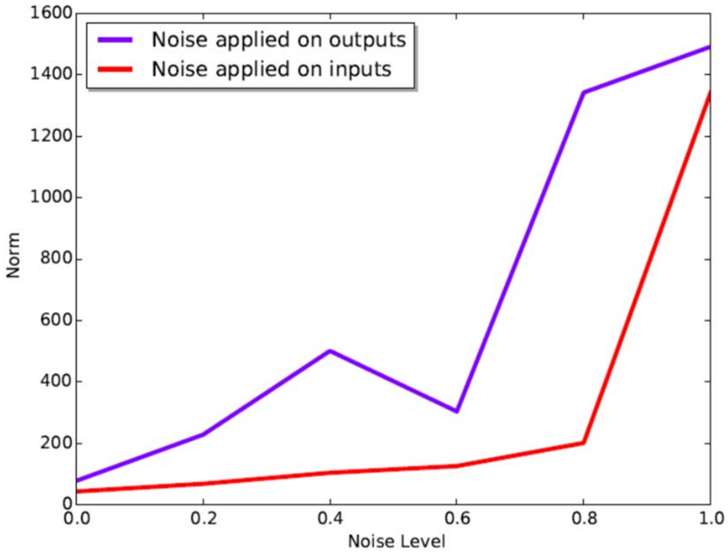
Other works

DEEP NETS DON'T LEARN VIA MEMORIZATION

David Krueger, …, Aaron Courville

ICLR 2017 workshops

"DNNs trained on real data examples learn simpler functions than when trained with noise data, as measured by the sharpness of the loss function at convergence."



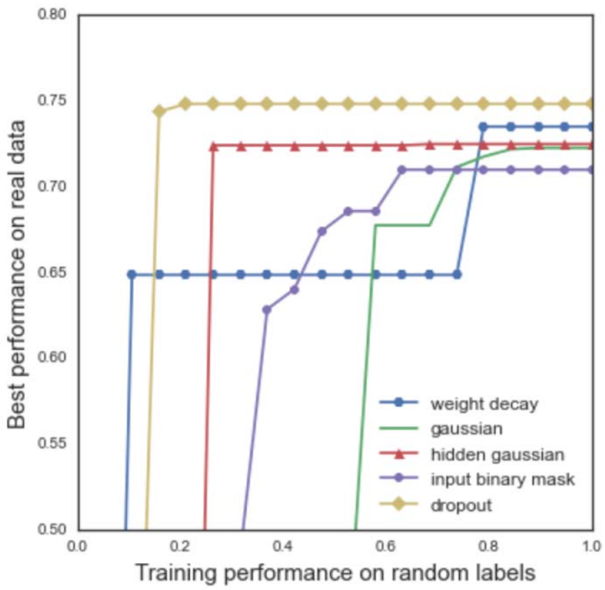
Other works

DEEP NETS DON’T LEARN VIA MEMORIZATION

David Krueger, …, Aaron Courville

ICLR 2017 workshops

“for appropriately tuned explicit regularization, e.g. dropout, we can degrade DNN training performance on noise datasets without compromising generalization on real data”



Other works

A Closer Look at Memorization in Deep Networks

ICML 2017

Other works

- On Generalization and Regularization in Deep Learning (explaining related topics on statistical learning theory in more details)

- **Behnam Neyshabour** et al. Exploring Generalization in Deep Learning (a closer look at different measures which can explain generalization of deep nets)

- Towards Understanding Generalization of Deep Learning: Perspective of Loss Landscapes (studies the difference of good and bad local minima by comparing the loss surface)

- Deep Learning is Robust to Massive Label Noise (closer look on training from noisy datasets. On MNIST they find that accuracy of above 90 percent is still attainable even when the dataset has been diluted with 100 noisy examples for each clean example.)

- High-dimensional dynamics of generalization error in neural networks (studies the dynamics of over-parametrized deep networks when trained using gradient descent)

- **Naftali Tishby**, Opening the Black Box of Deep Neural Networks via Information (the goal of the network is to optimize the Information Bottleneck (IB) tradeoff between compression and prediction, successively, for each layer)

- **Behnam Neyshabour** et al. In Search Of The Real Inductive Bias : On The Role Of Implicit Regularization In Deep Learning (directly explaining why one should go beyond the deep network parameter size to explain complexity control mechanisms of deep nets)

Final word

So, we need to “rethink” generalization when it comes to deep networks

That probably boils down to

understanding *how deep networks prefer simpler solutions* while capable of memorizing more than simple patterns