Understanding deep learning requires rethinking generalization

Chiyuan Zhang*
Massachusetts Institute of Technology
chiyuan@mit.edu

Samy Bengio
Google Brain
bengio@google.com

Moritz Hardt
Google Brain
mrtz@google.com

Benjamin Recht†
University of California, Berkeley
brecht@berkeley.edu

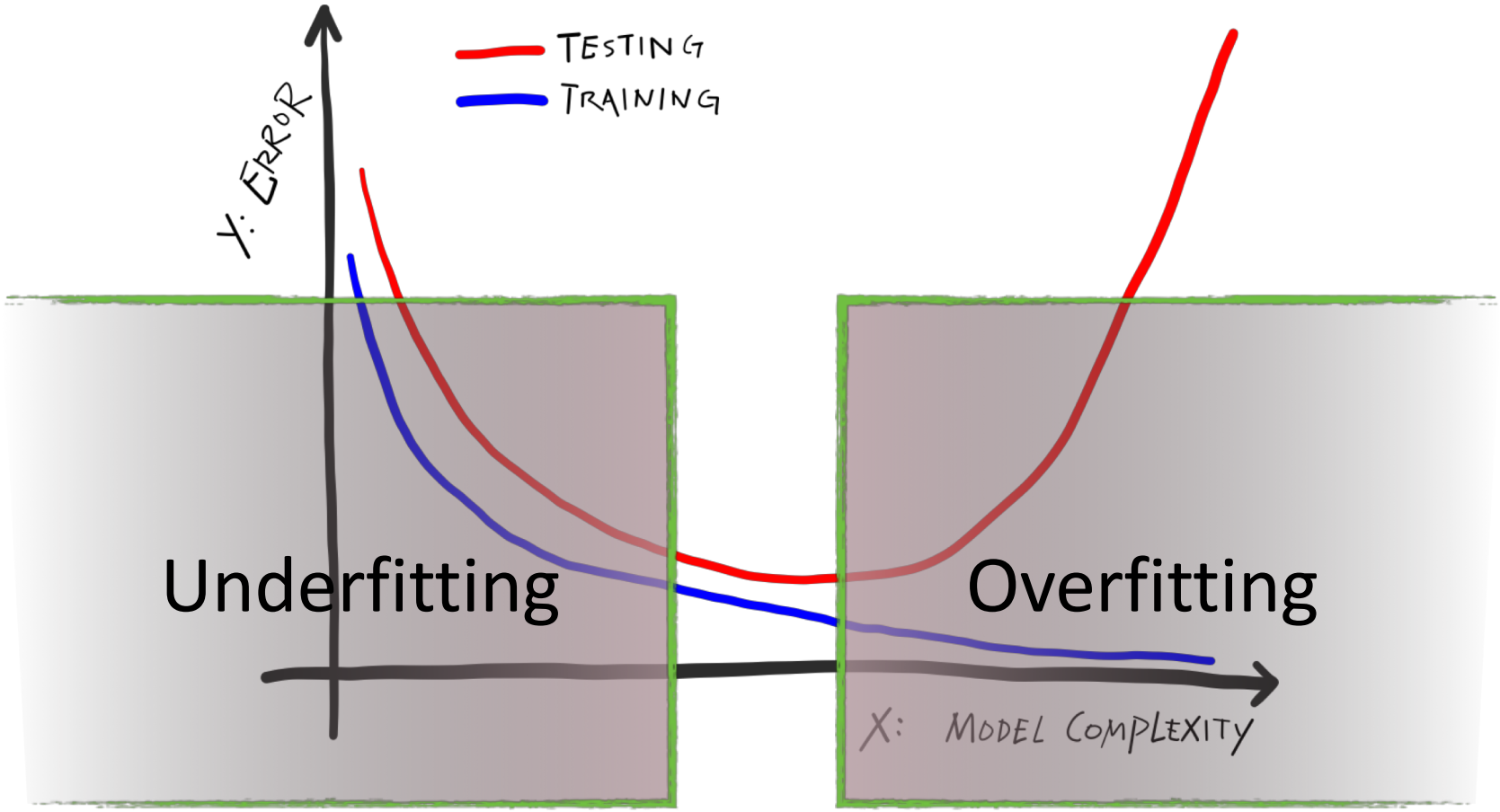
Oriol Vinyals
Google DeepMind
vinyals@google.com

ICLR Best Paper Award, 2017

Summary…

- Generalization error is the difference between training error and test error.
- The author raised a question “What is it that distinguishes neural networks that generalize well from those that don’t”.
- “In this work, we problematize the traditional view of generalization by showing that it is incapable of distinguishing between different neural networks that have radically different generalization performances.”

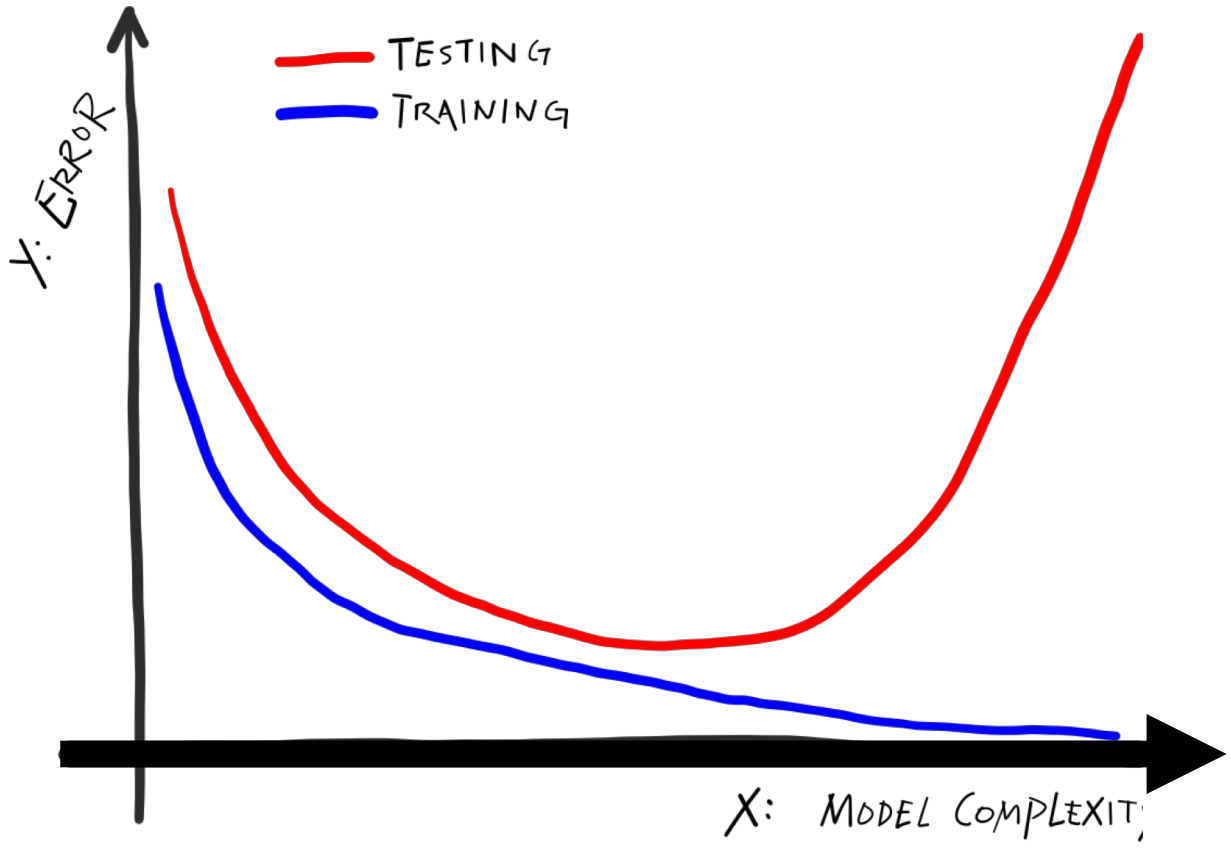
Model Selection



Deep learning

CIFAR-10	# train: 50,000
Inception	1,649,402
Alexnet	1,387,786
MLP 1x512	1,209,866
ImageNet	# train: ~1,200,000
Inception V4	42,681,353
Alexnet	61,100,840
Resnet-{18;152}	11,689,512; 60,192,808
VGG-{11;19}	132,863,336; 143,667,240

Bias — Variance



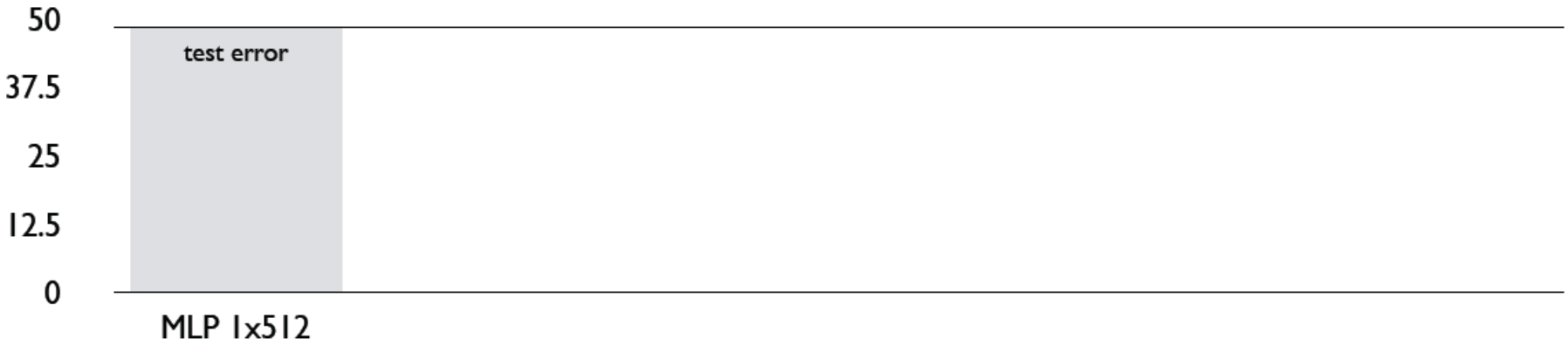
Deep Learning

$$\frac{\text{Parameter Count}}{\text{Num Training Samples}}$$

Parameter Count / Num Training Samples

MLP 1x512

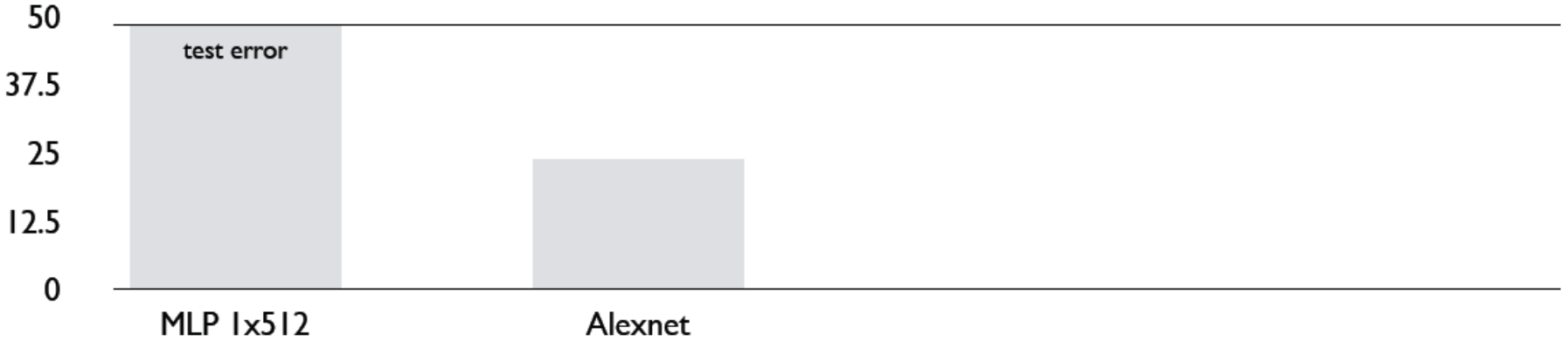
p/n: 24



Parameter Count / Num Training Samples

Alexnet
p/n: 28

MLP 1x512
p/n: 24

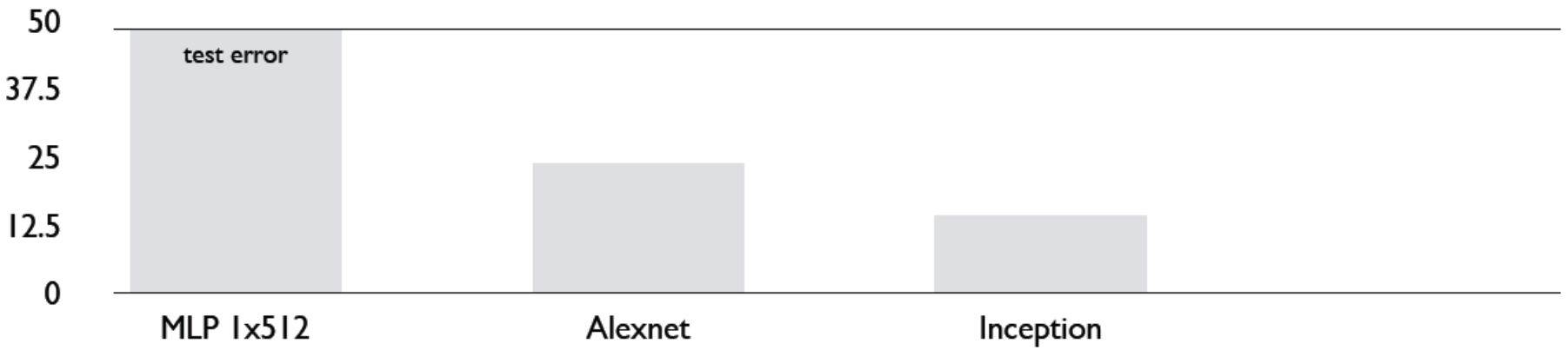


Parameter Count / Num Training Samples

Inception
p/n: 33

Alexnet
p/n: 28

MLP 1x512
p/n: 24



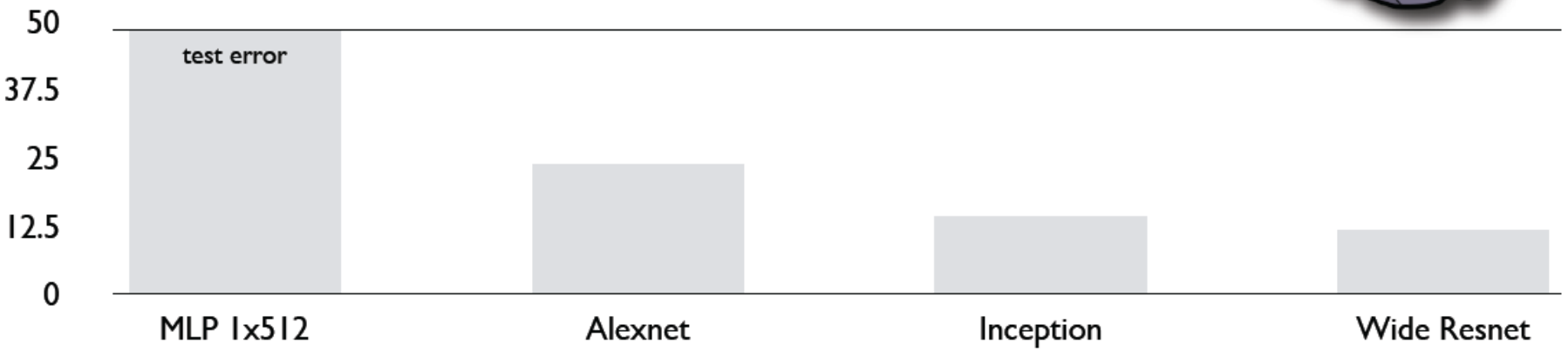
Parameter Count / Num Training Samples

MLP 1x512
p/n: 24

Alexnet
p/n: 28

Inception
p/n: 33

Wide Resnet
p/n: 179

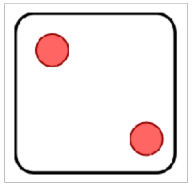


Randomization Test

Deep Neural Networks easily fit random labels.

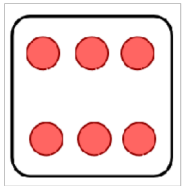
Random Label Dataset

Dog

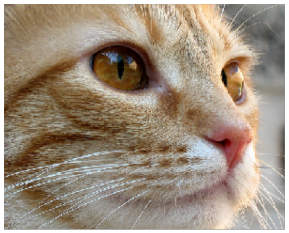


Cat

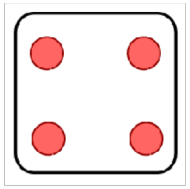
Flower



Dog

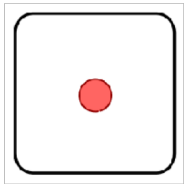


Cat



Bus

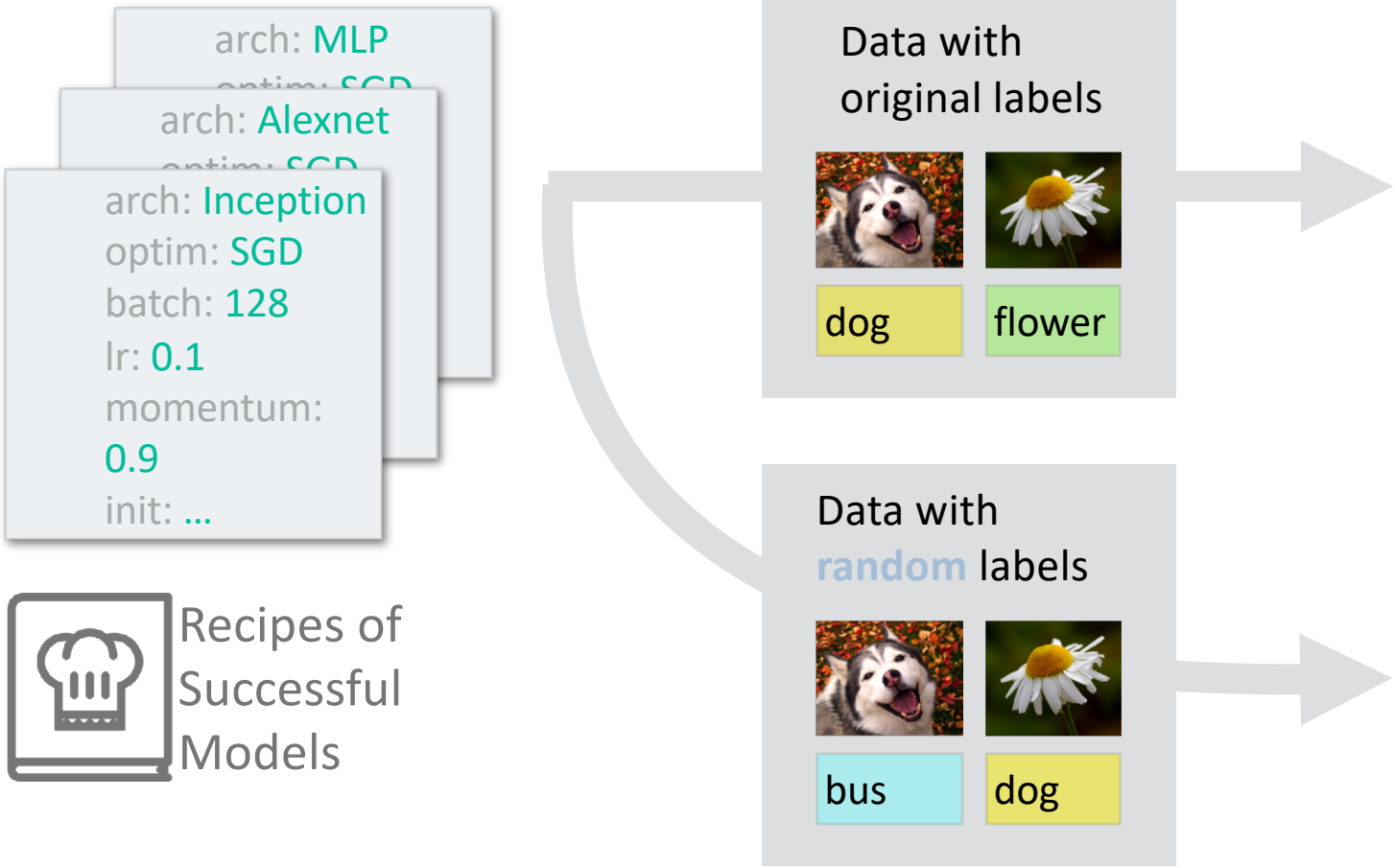
Flower



Bird

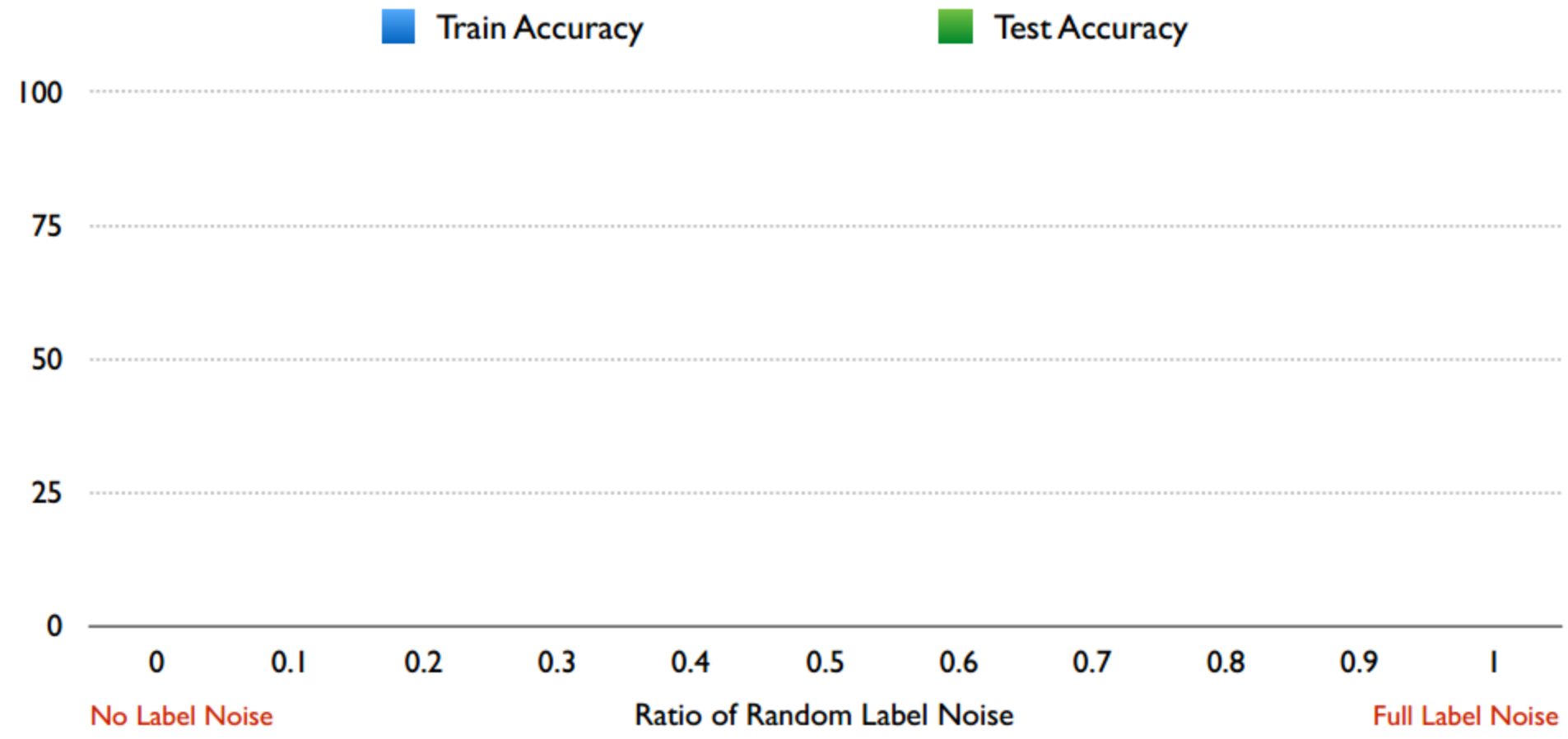
⋮

Randomization Test

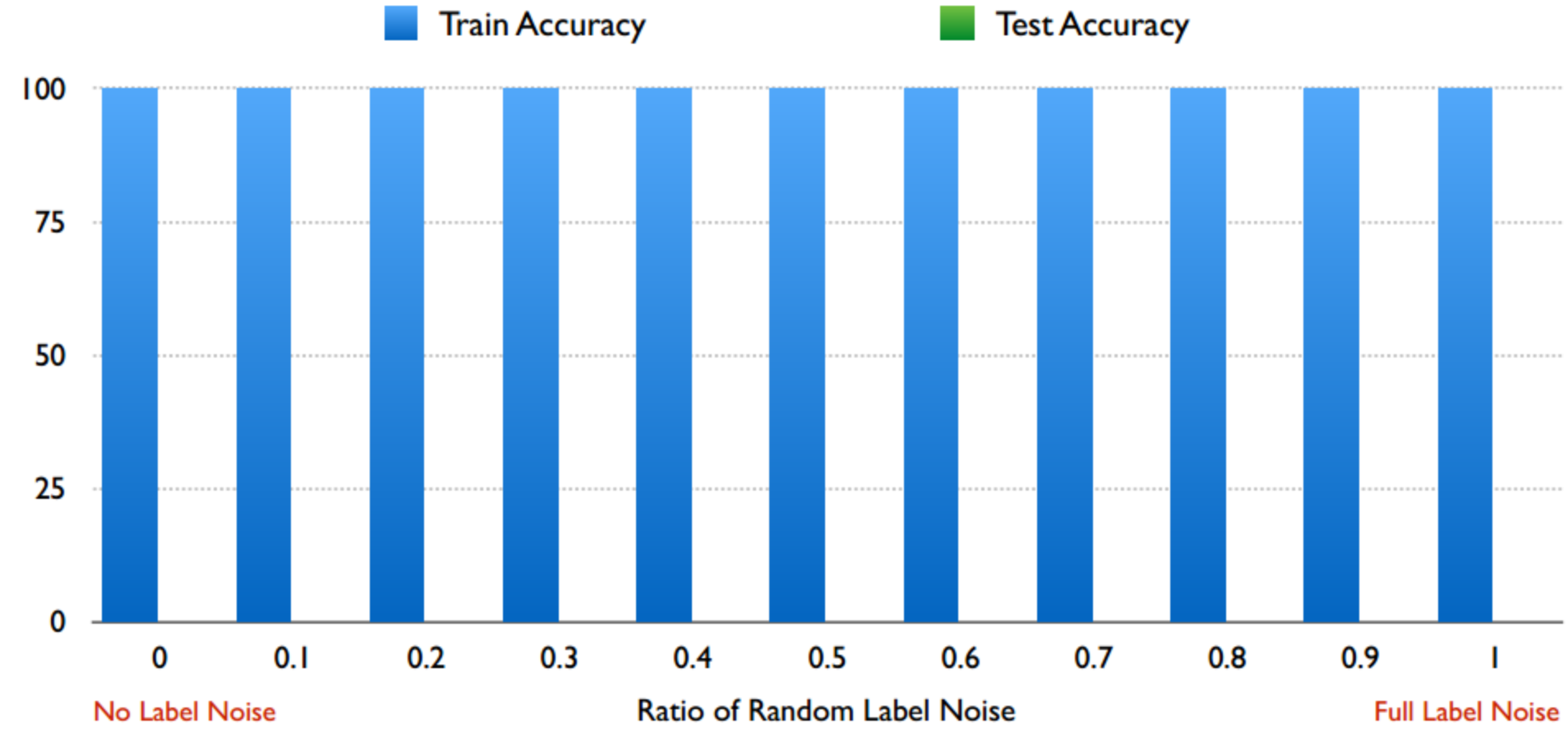


https://www.tensorflow.org/
http://mxnet.io/

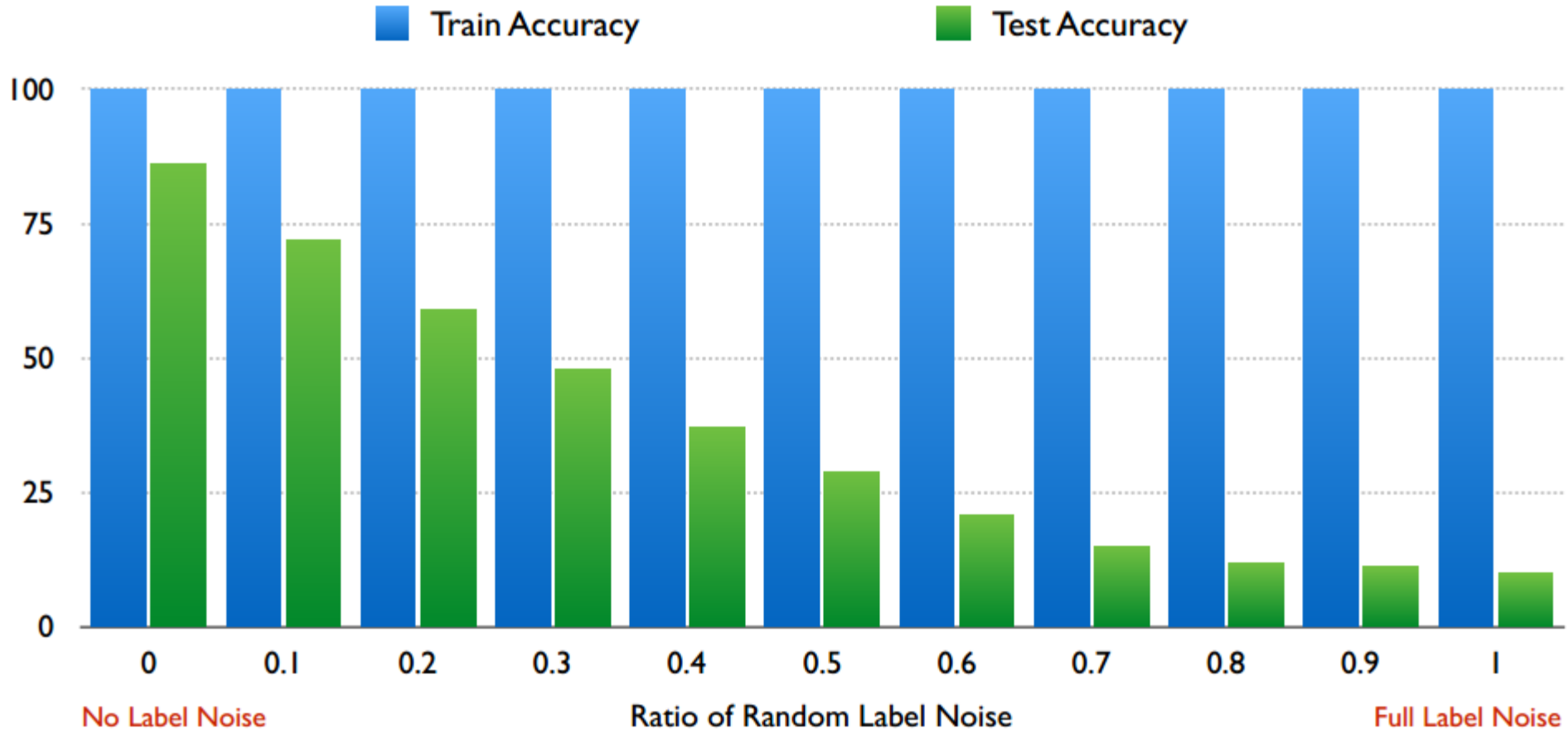
Randomization Test



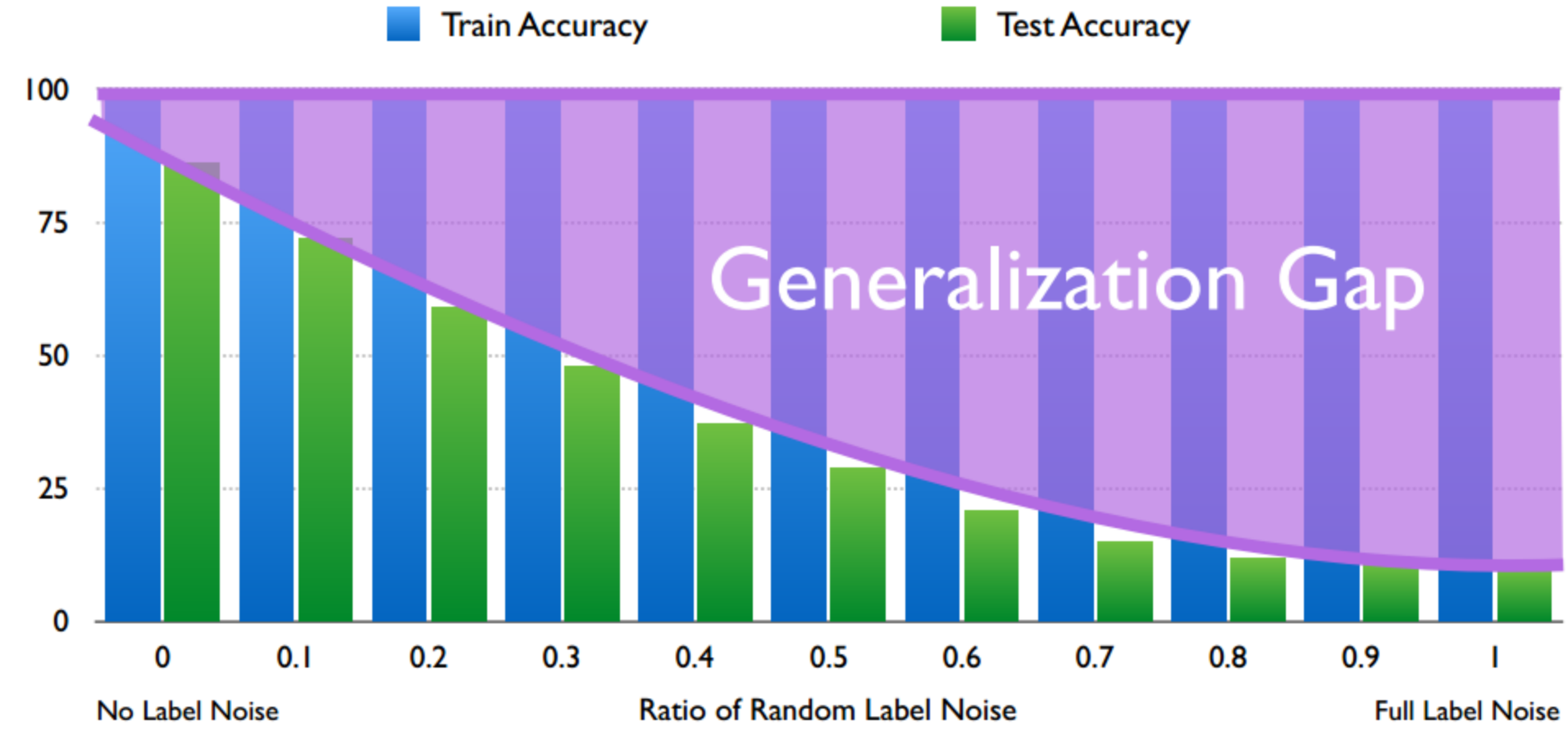
Randomization Test



Randomization Test



Randomization Test



Implication of Randomization Test

- Rademacher complexity measures ability of model to fit random ±1 binary label assignments. Because NNs fit the random labels perfect, R~1. But this is upper bound for Rademacher complexity.
- Uniform Stability measures how sensitive the algorithm is to the replacement of a single example. This does not take into account specifics of the data or the distribution of the labels.

Randomization Test

Deep Neural Networks easily fit random labels.

Regularizers in Deep Learning

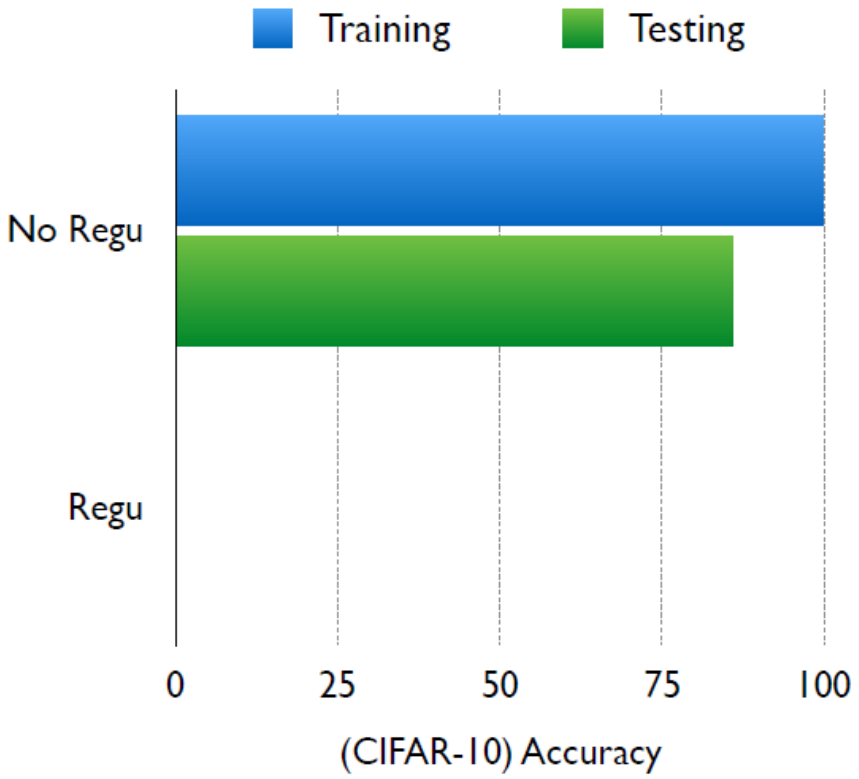
- Data augmentation: domain-specific transformations

- Weight decay: l2-regularizer on weights

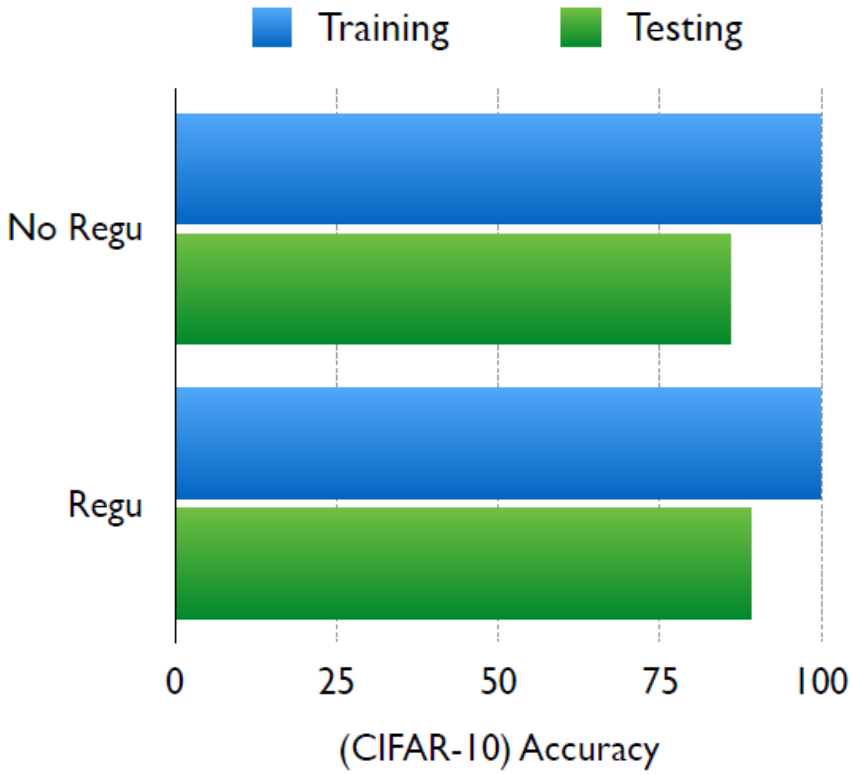
- Dropout*: randomly mask out responses

* Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. JMLR, 15(1):1929–1958, 2014.

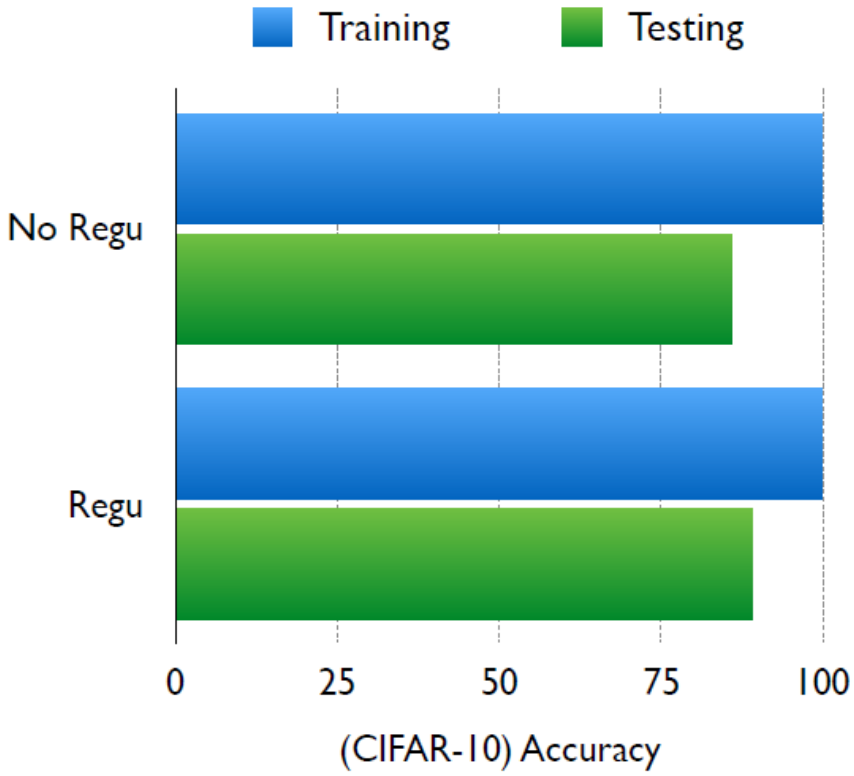
Fitting Natural Label with Regularizers

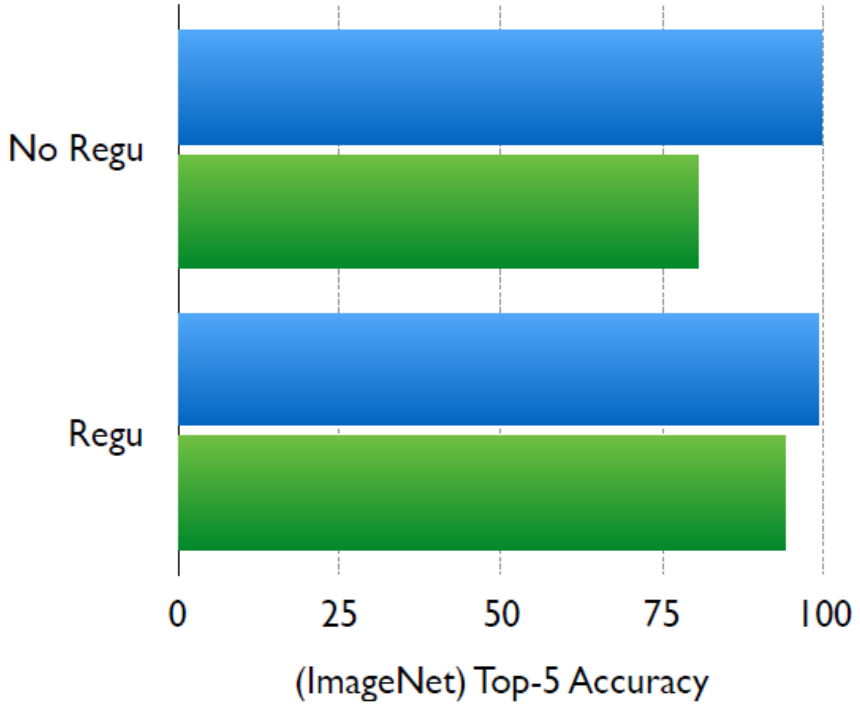


Fitting Natural Label with Regularizers



Fitting Natural Label with Regularizers





Fitting Random Label with Regularizers

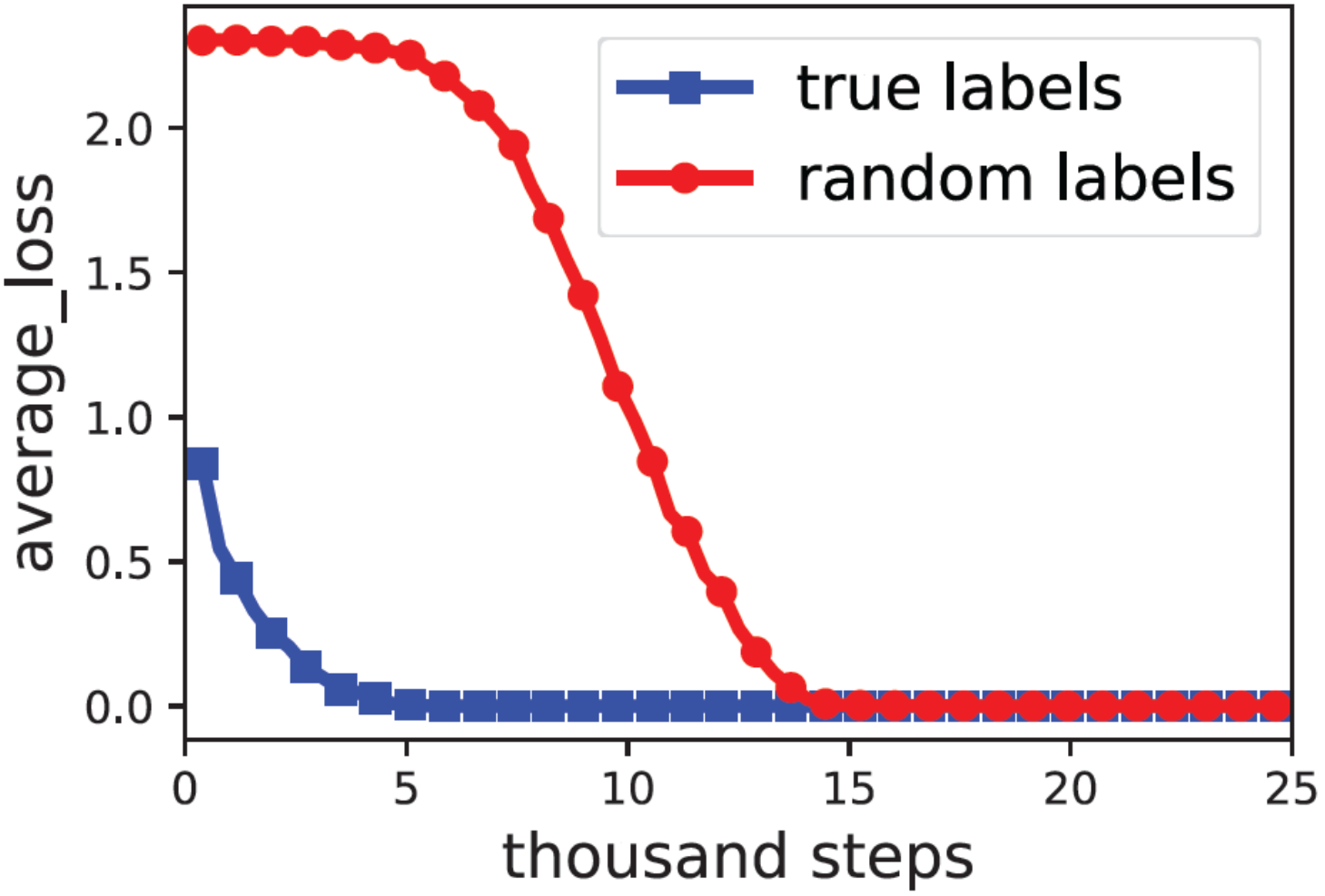
Regularizer	Model	Training Accuracy
Weight decay	Inception	100%
	Alexnet	Failed to converge
	MLP 1x512	99.21%
Crop Augmentation*	Inception	99.93%

IMAGENET

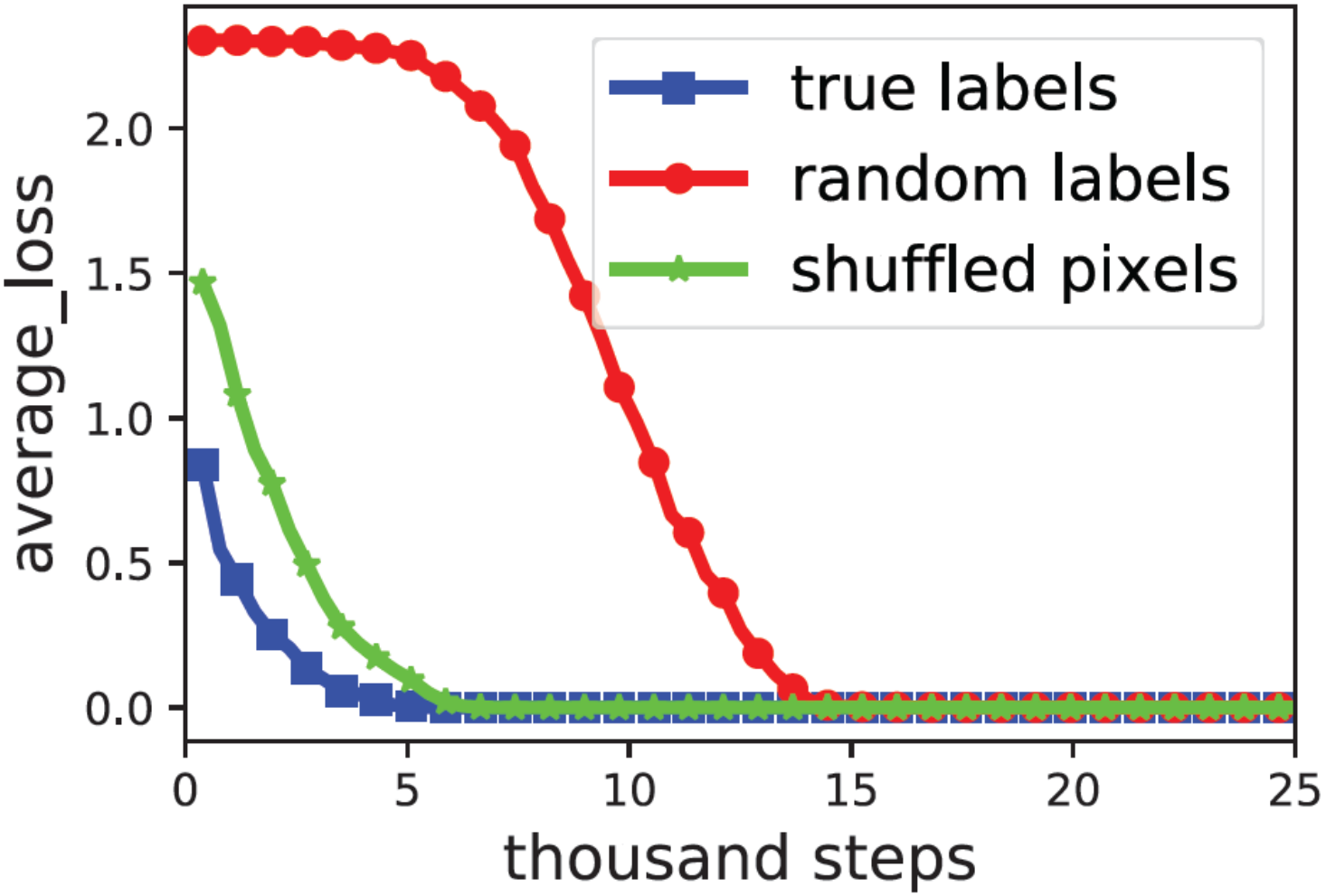
Regularizer	Model	Training top-5
Dropout	Inception V3	96.15%
Dropout + Weight decay		97.95%

* We need to tune the hyperparams a bit and run for more epochs for this to converge, see paper for details.

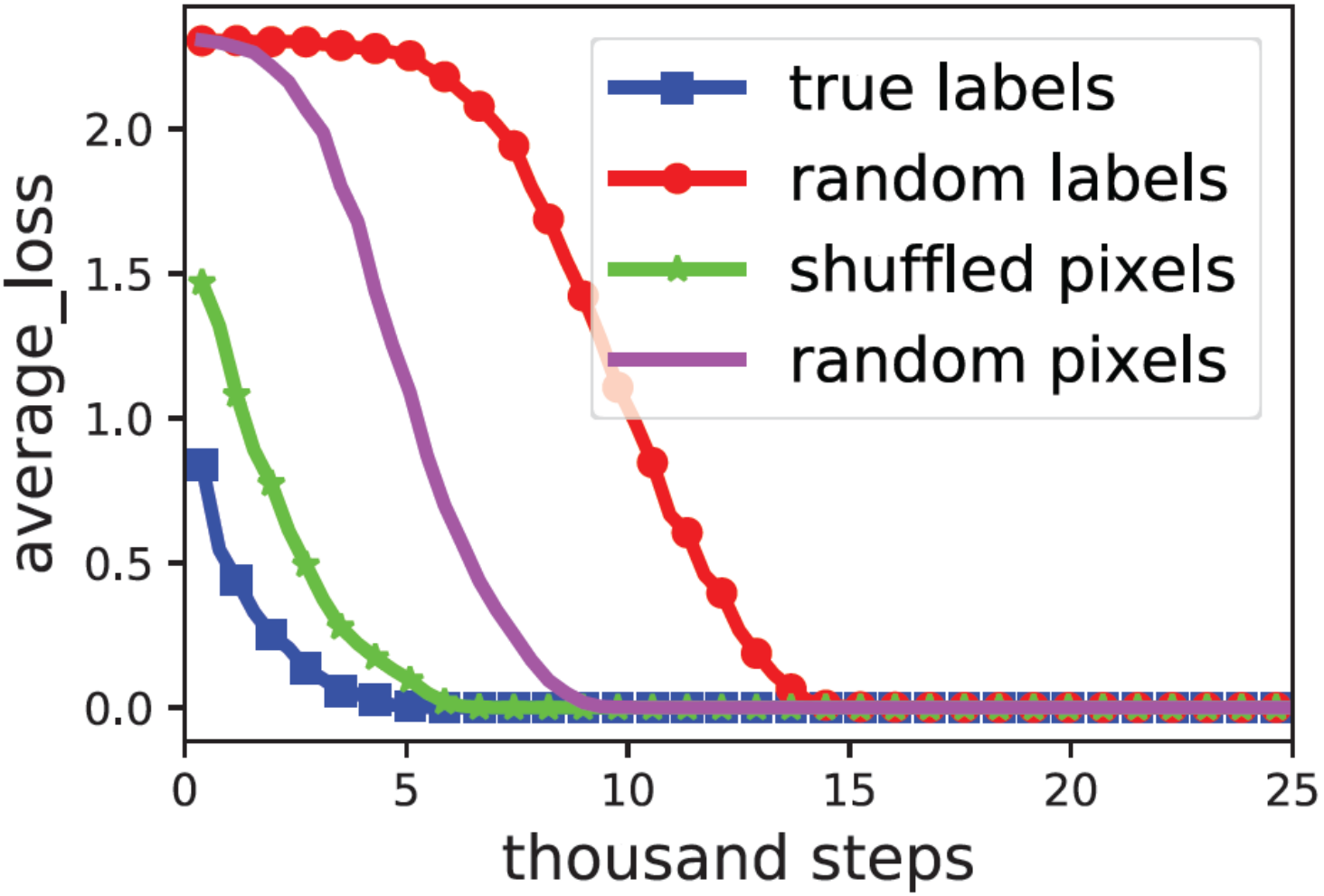
SGD fits Random Labels



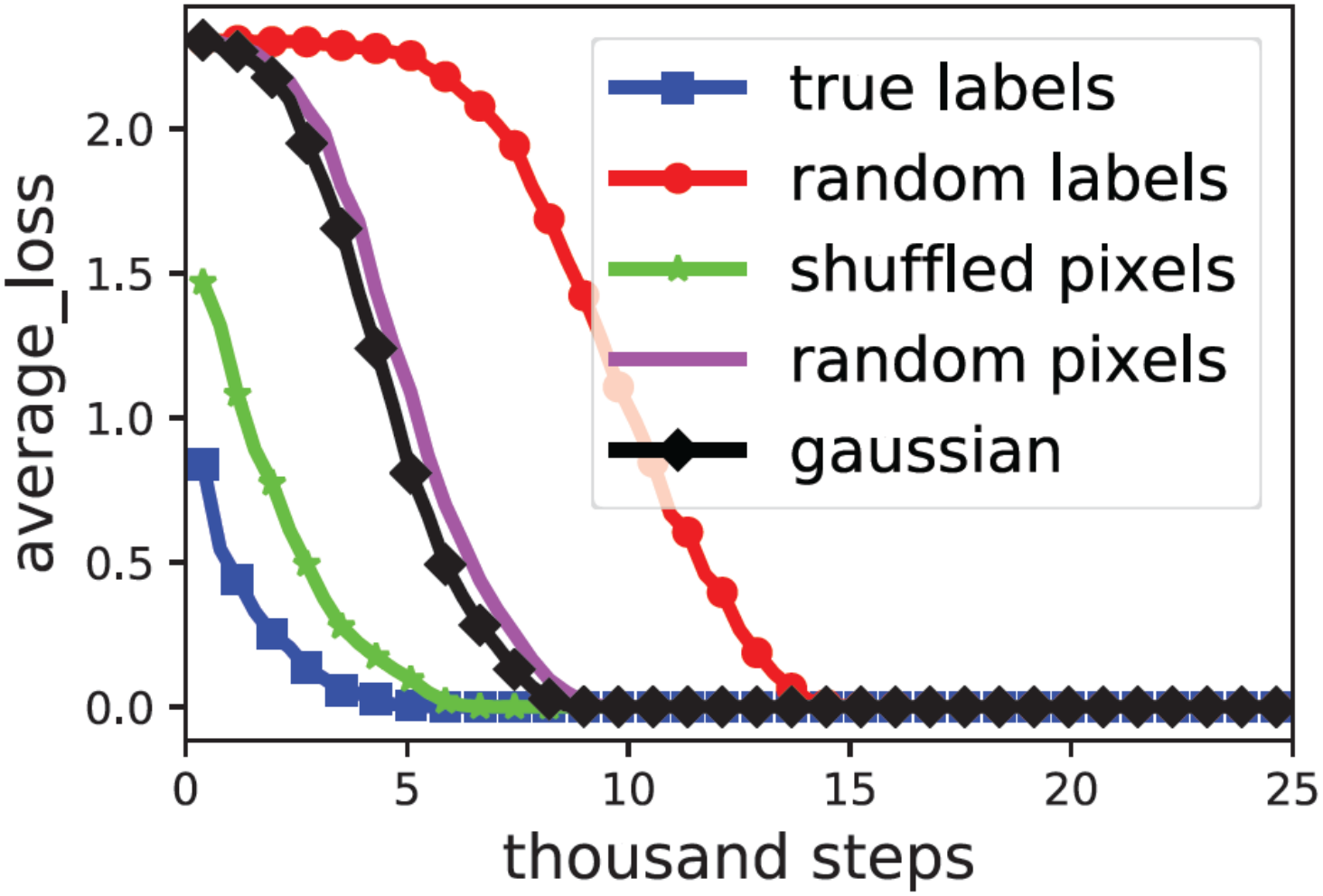
SGD fits Random Labels



SGD fits Random Labels



SGD fits Random Labels



Regularizers

Regularizers could help to improve the generalization performance, but it is unlikely that the regularizers are the fundamental reason for generalization.

Fitting Random Labels

Optimization is “easy” for deep learning.

Source of difficulty for optimization and generalization are not necessarily correlated.

Conclusions

- Deep neural networks easily fit random labels.
- Explicit regularization may improve generalization performance, but is neither necessary nor by itself sufficient for controlling generalization error.
- Other formal measures of complexity for the models / algorithms / data distributions are needed to precisely explain the over-parameterized regime.

Slices adapted from http://pluskid.org/publications/