Reproducible Study of Training and Generalization Performance

MATH63800 Final Project

XIAO Jiashun, WANG Ya, LIU Yiyuan, YU Tingyu

Content

- Introduction
- Background
- Experiment
- Conclusion

Introduction

What we do?

- Reproduce the best award paper in ICLR2017:
  Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals, Understanding deep learning requires rethinking generalization.
- Compare our results with the paper and try to explain.

What they do?

- Try to find:
  What is it then that distinguishes neural networks that generalize well from those that don't?
- They conduct:
  - Randomization tests
  - Turn on regularization
- Their conclusion:

  It is incapable of distinguishing between different neural networks that have radically different generalization performance.

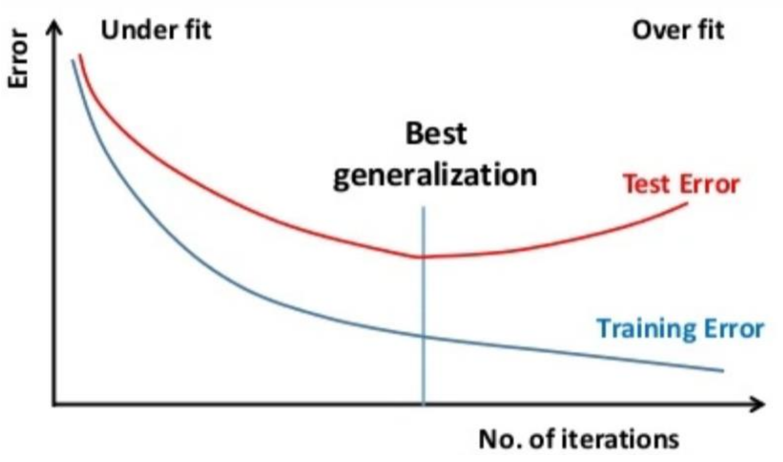
Background

Model capacity

The flexibility of the networks and all the different types of inputs that it could fit to.

Generalization error

Difference between training error and test error.



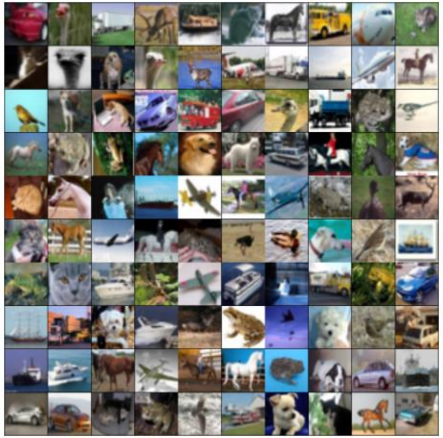
Experiment

Setup:

Dataset

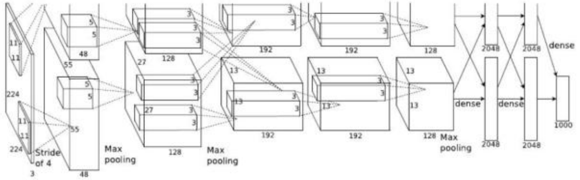
FashionMNIST

CIFAR10

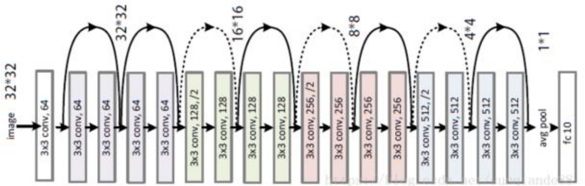


Model architecture

Alexnet



Resnet18



Experiment

Randomization tests

*True labels *Random labels *Random pixels

label

Dog

Cat

Monkey

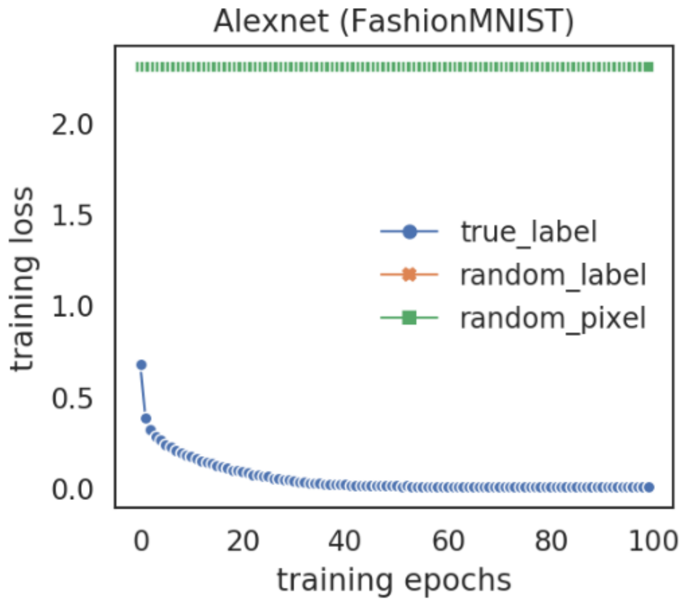
Bear

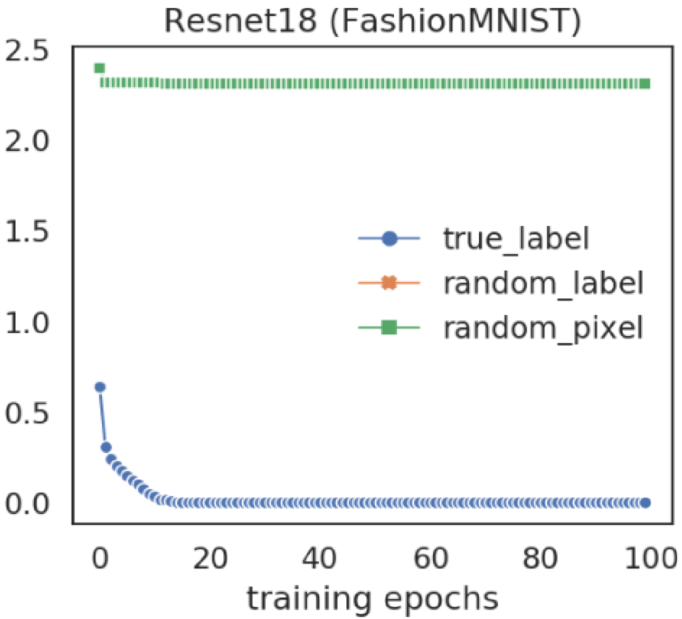
Panda

Human

Experiment

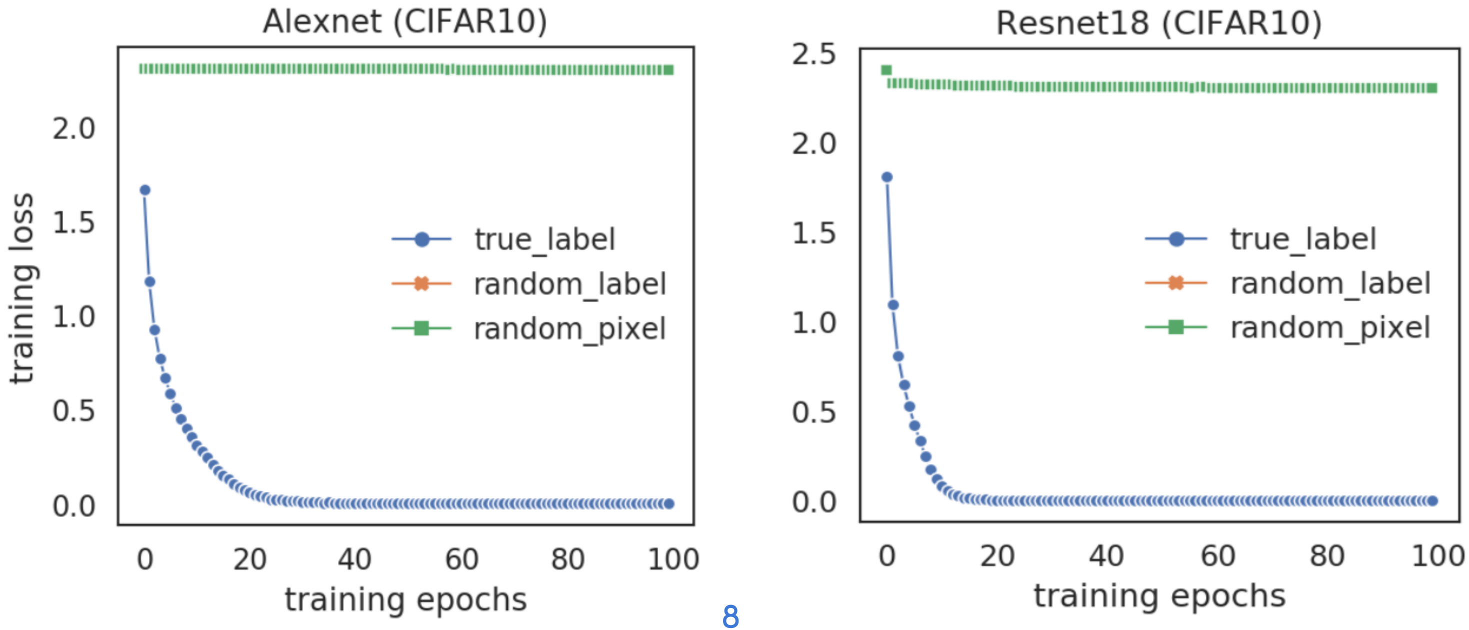
Results of randomization tests -- FashionMNIST





Experiment

Results of randomization tests – CIFAR10



Experiment

Implication of randomization tests:

- The true labels are able to fit to the data pretty quickly and converge to 0 training error.
- For the random labels, the average loss also could converge to zero as the same as the true labels.
- For random pixels, training error doesn't seem to converge to zero. One possible reason is that the number of epochs is not enough in our experiments, another reason is the incorrect hyperparameters that we set in our models.

→ The model capacity of neural networks is sufficient for memorizing the entire dataset.

Experiment

Explicit regularization test:

Weight decay: equivalent to l_2 regularizer on the weights.

standard weight update:

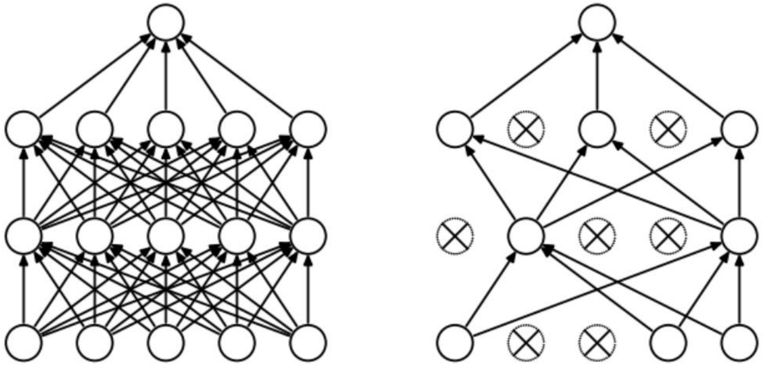
$$w_{t+1} \leftarrow w_t - \eta_t \frac{\partial L(w)}{\partial w_t}$$

$$\mathrm{L(w)} = \mathrm{L}_0(w) + \frac{\lambda}{2} w^{2*}$$

New weight update:

$$\mathrm{w}_{t+1} \leftarrow w_t - \eta_t \frac{\partial L_0(w)}{\partial w_t} - \eta_t \lambda w_t$$

Dropout: mask out each element of a layer output randomly with a given drop probability.

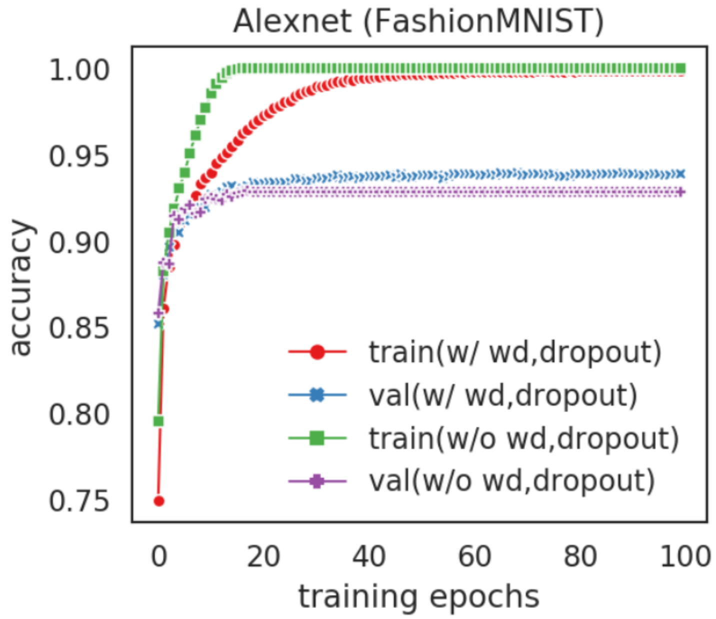


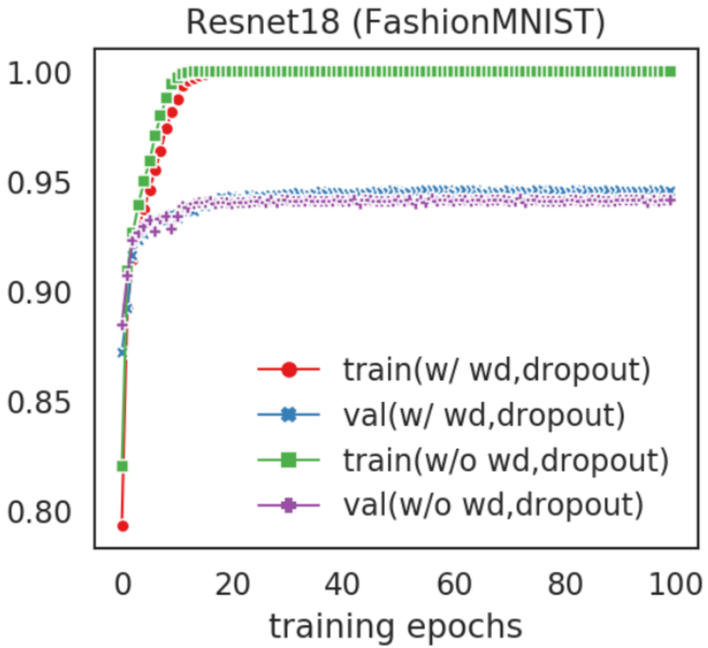
(a) Standard Neural Net

(b) After applying dropout.

Experiment

Results of explicit regularization test:





Experiment

Implication of explicit regularization test:

- Both explicit generalization techniques help to improve the performance.
- However, the model can also generalize well without regularizers.
- Changing the model architecture from Alexnet to Resnet18 can achieve better gains that explicit regularizer.

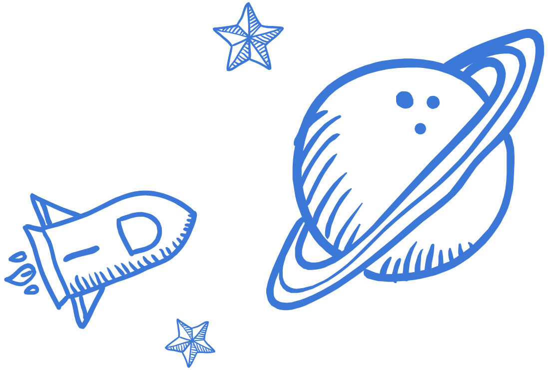
→ As what states in the original paper, explicit regularization is neither nor by itself sufficient for controlling generalization error.

Conclusion

We get similar results with the original paper.

- Neural network model could fit the random labels easily and perfectly.
- Explicit regularizers do not count as a fundamental phase change in the generalization capacity of deep nets.
- It is incapable of answering the question at the beginning of the paper.

  i.e. We cannot distinguish neural networks that generalize well from those that don't.



THANK YOU