Reproducible Study of Training and Generalization Performance

XIAO Jiashun, YU Tingyu, LIU Yiyuan, WANG Ya

MATH6380O Final Project
HKUST, Department of Mathematics

Abstract

The paper *Understanding deep learning requires rethinking generalization*[6] was selected as one of the three best award papers in ICLR 2017, which still receives widespread concern. We want to reproduce some key experiments and verify whether we can get same results in this paper, such as randomization tests and explicit regularization experiment. The new point is that we conduct experiments on different dataset using different model architectures. After then, we analysis our results and compare them with original paper.

Introduction

In deep articial neural networks, the number of trainable model parameters always exceeds the number of training samples, which always results in overfitting problems in statistical learning theory. However, some neural network models exhibit small difference between "training error" and "test error", while many others generalize poorly. The original paper aims to find out the reason why different models perform differently.

Dataset

Fashion-MNIST:
Fashion-MNIST is dataset of Zalando's article images. It consists of a training set of 60000 examples and a test set of 10000 examples. Each example is a 28*28 grayscale image, associated with a label from 10 classes.
CIFRA10:
CIFAR10 is labeled subsets of the 80 million tiny images dataset, which is collected by Alex Krizhevsky, Vinod and Geoffrey Hinton. It contains 60000 32*32 colour images in 10 classes, with 50000 training images and 10000 test images.

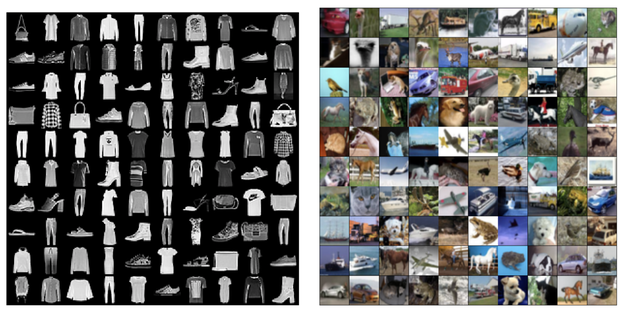

Figure 1: Schematic diagram of the dataset

Network Setting

We apply two state-of-the-art neural networks, the so called Alexnet[2] and Resnet 18[1] to classification.

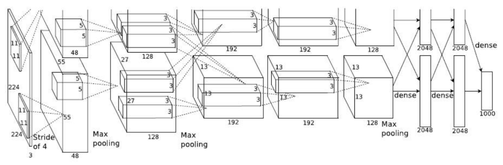

Figure 2: Alexnet - eight layers convolutional network

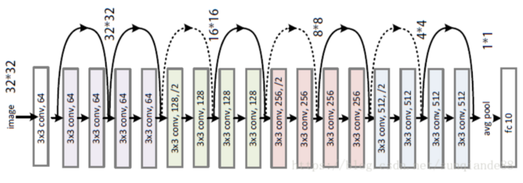

Figure 3: Resnet 18 - a residual network with 18 layers

For all experiments on both of the dataset, we train using SGD with a momentum parameter of 0.9. An initial learning rate of 0.1 for Alexnet and 0.01 for Resnet 18 are used, with a decay factor of 0.00001 and a dropout probability of 0.5 per training epoch.

Model Capacity and Regularization

Model Capacity:
Model capacity means the flexibility of the networks and all the different types of inputs that it could fit to.

- Universial approximation theorem states that simple neural networks can represent a wide variety of interesting functions when given appropriate parameters; however, it does not touch upon the algorithmic learnability of those parameters.
- VC-dimension[5] of a model f is the maximum number of points that can be arranged so that f shatters them. But this measure of capacity is not useful for neural network when the number of parameter exceeds the number of data points.

Regularization:
Regularizers play an important role in theory and practice to mitigate over-fitting in the regime when there are more parameters than the data points. It could be divided into explicit regularizers and implicit regularizers. In our experiments, we add two explicit regularizers to the neural network models, including weight decay and dropout.

- Weight Decay[3]: equivalent to a L_2 regularizer on the weights, which forces the weights to be small by imposing a L_2-penalty to the loss function and makes all the weights close to lower values.
- Dropout[4]: mask out each element of a layer output randomly with a given dropout probability. It forces the neural network not to rely on the individual neurons by randomly drop neurons from layers in the network.

Randomization tests

In this part, we conduct experiments on CIFRA10 and Fashion-MNIST using Alexnet and Resnet18 architectures.
Experiments include following modification of the labels and input images:

- True labels: without modification.
- Random labels: arrange all the images to the labels randomly.
- Random pixels: every image data has been arranged to its own permutation independently.

Result:

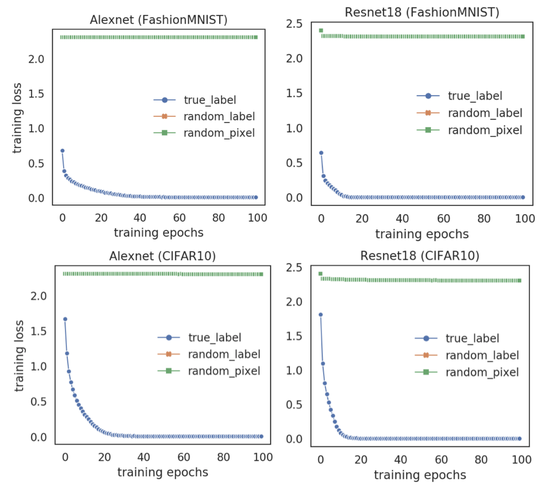

Figure 4: Fitting True labels, random labels and random pixels on Fashion-MNIST and CIFRA10;

Figure 4 shows the training loss of various experiment settings decaying with the number of epochs from 0 to 100. We can get the facts below:

- 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 could also converge to zero as the same as the true labels.
- For random pixels, training error does not 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 hyper-parameters setting in our models.

Implication:
An interesting discovery about the results is that even when we completely destroy the relationship between images and labels, we can still perfectly fit them. It indicates that deep neural networks easily fit random labels and their model capacity is sufficient for memorizing the entire dataset.

Explicit regularization

In this part, we add two different kinds of explicit regularizers (Weight Decay and Dropout) to explore whether they can control generalization error.
Results:

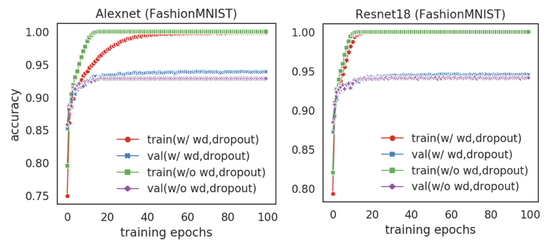

Figure 5: Effects of explicit regularizers on generalization performance.

Figure 5 gives us following information:

- Both techniques exactly help to improve the generalization performance.
- The models still generalize very well even without regularizers, which is the interesting point.
- Changing the model architecture from Alexnet to Resnet18 can achieve bigger gains than turning on regularizers.

Implication:
All of these information states that regularizers cannot count as a fundamental phase change in the generalization capability of deep nets.

Conclusions and Discussion

After reproducing experiments of the original paper, we get some similar results, such as the neural network model could fit to the random labels perfectly and turning on the explicit regularizers could improve the generalization performance. However, when we do the randomization tests corresponding to the random pixels, the training error doesn't show any decreasing trend along the number of epochs. This different result is either caused by the lack of iteration numbers, or the incorrect parameters we set. More importantly, neither the properties of the model family nor the regularization techniques can explain the small generalization error. It is incapable of answering the question at the beginning of the paper, which means we still cannot distinguish neural networks that generalize well from those that don't.

Acknowledge

Xiao Jiashun: Explore and implement models to randomization tests on CIFRA10 and design poster.
Yu Tingyu: Implement Resnet18 to randomization test on dataset and design slides.
Liu Yiyuan: Explore and implement Alexnet to tests on Fashion-MNIST and design poster.
Wang Ya: Explore and implement models to explicit regularization and record presentation.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[3] Anders Krogh and John A Hertz. A simple weight decay can improve generalization. In *Advances in neural information processing systems*, pages 950–957, 1992.

[4] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[5] A. Ya. Chervonenkis V. N. Vapnik. The uniform convergence of frequencies of the appearance of events to their probabilities. *Dokl. Akad. Nauk SSSR*, 181:781–783.

[6] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.