Rethinking generalization and Non-overfitting puzzle

Liu Ping, Liu Yixuan, Rong Yi, Zhu Chang
Department of Mathematics
Hongkong University of Science and Technology

Abstract

We re-implement some experiments to illustrate the effective capacity and non-overfitting puzzle of neural network. We also find an interesting property of neural network in the experiment. Besides, we guess the data set may be the reason of this non-overfitting puzzle and give an experiment to support our guess.

1 Effective Capacity of neural network

We make some experiments to show that no matter what a complex and weird data set we have, we can achieve zero training loss if we have enough parameters. This means the capacity of neural network is effective. Besides, we observe some phenomenon which indicates interesting property of neural network.

1.1 Experiments1

We run many experiments with the following modifications of the labels and input images:

- True labels: the original dataset without modification.
- Partially corrupted labels: independently with probability p, the label of each image is corrupted as a uniform random class.
- Random labels: all the labels are replaced with random ones.
- Shuffled pixels: a random permutation of the pixels is chosen and then the same permutation is applied to all the images in both training and test set.
- Random pixels: a different random permutation is applied to each image independently.
- Gaussian: a Gaussian distribution (with matching mean and variance to the original image dataset) is used to generate random pixels for each image.
- Mosaic: each training image is replaced by a 'mosaic' image with probability 0.9, all mosaic images are generated by one Gaussian distribution.

Figure 1 shows the learning curves of the wideResnet model on the CIFAR10 dataset under various settings. We expect the objective function to take longer to start decreasing on random labels because initially the label assignments for every training sample is uncorrelated. And we observe that the training loss will always decay to zero regardless of how weird data we deal with. This illustrates the effective capacity of neural network.

From Figure 2 We observe a steady increasing of the convergence time as we increase the noise level. This shows that neural networks need more time to achieve overfitting for a more complex model.

As shown in Figure 2 & 3, mosaic dataset behaves better in both convergence time and test error than other noisy dataset, it seems that deep neural network can automatically 'learn' from those correct images and ignore those mosaic images. On the other hand, deep learning is not that robust to massive noisy labeled dataset.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

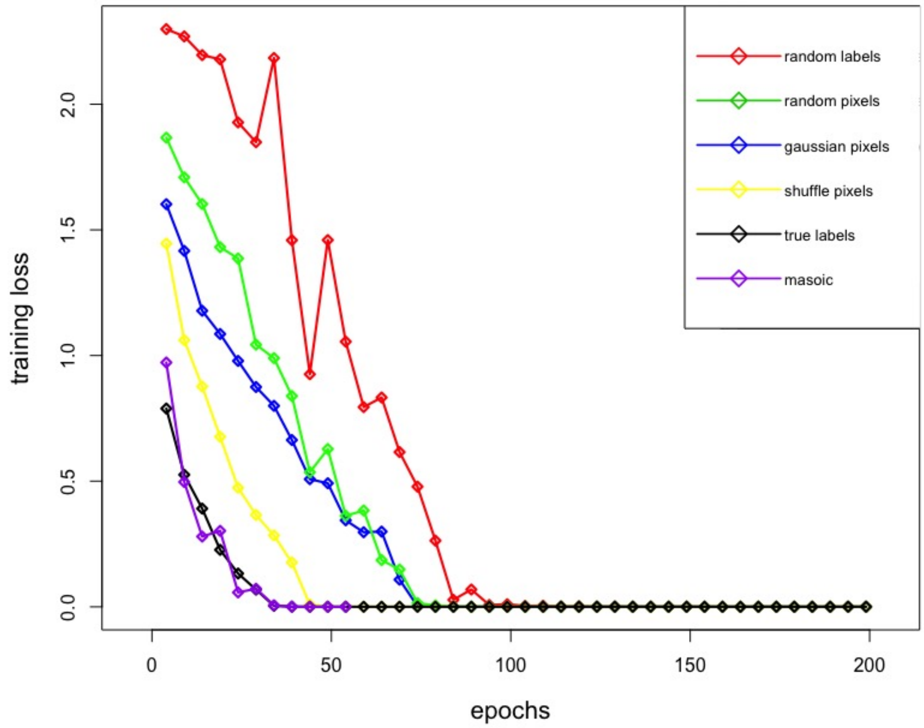

Figure 1: Fitting random labels and random pixels on CIFAR10. It shows the training loss of various experiment settings decaying with the training epochs.

2 Non-overfitting puzzle of Deep learning

A main puzzle of deep networks revolves around the absence of overfitting despite large overparametrization and despite the large capacity demonstrated by zero training error on randomly labeled data(shown in experiment1). We use some experiments to illustrate non-obverfitting behaviors of deep networks classification and overfitting in SVM classification problem. This gives a direct explanation of the non-overfitting puzzle in deep learning. Besides, we have some guess on the reason of this intriguing puzzle based on experiment.

2.1 Experiment2

We use MLP model deal with the CIFAR10 classification problem. We increase the parameters slowly and record the related loss and classification error in both training and test set.

As it shown by the Figure 4, we observe the overfitting phenomenon in test loss but non-overfitting phenomenon when viewing behavior of test error.

3 Possible reason of the non-overfitting puzzle

3.1 Overfitting experiments

We design an overfitting classification experiment serving to explain overfitting and non-overfitting which finally gives some intuition to understand what may cause non-overfitting puzzle. We use SVC to do a two set classification job.

Figure 5 gives a direct illustration of overfitting.

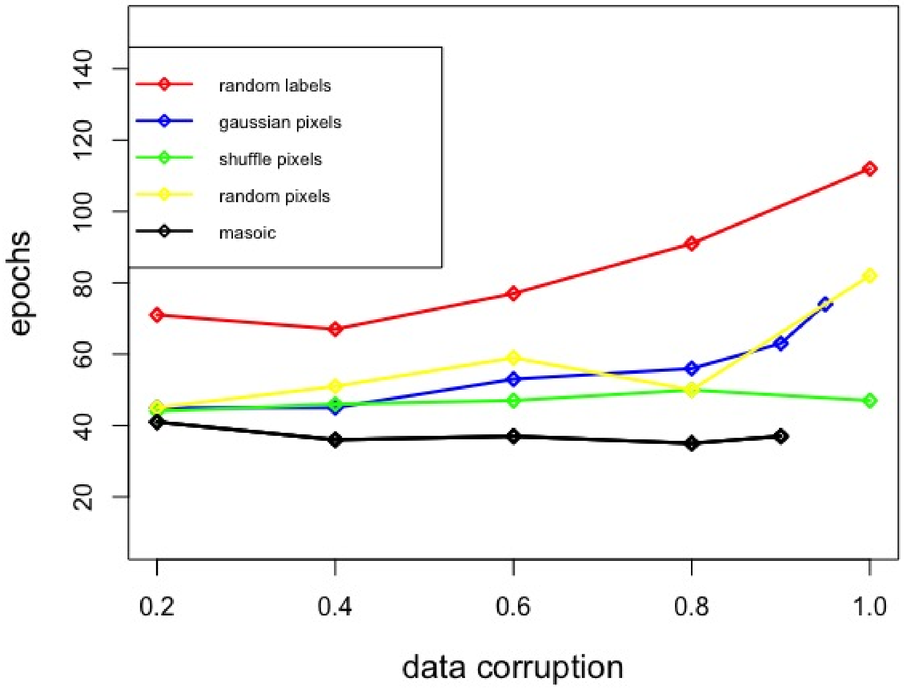

Figure 2: Relative convergence epoch with different data corruption ratio

3.2 Possible reason

We use Figure 6 to illustrate some reason and understanding we guess for the non-overfitting phenomenon of deep learning.

When $\gamma > 10$, from the Figure 5 we see that, there is an overfitting phenomenon. But this time, from Figure 6 we can not see the overfitting phenomenon in the training and test error view. This example gives us an indication that the overfitting puzzle in deep learning may caused by data set. That is, although from the picture we may see the overfitting phenomenon but we can not see it in error view. Because there is some similarity between training and test set which prevents the test error being worse when overfitting happened.

Acknowledgments

Thanks Guo Feifei for some help in coding.

References

[1] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht& Oriol Vinyals (2016) Understanding deep learning requires rethinking generalization *arXiv:1611.03530, 2016.*

[2]T. Poggio, K. Kawaguchi, Q. Liao, B. Miranda, L. Rosasco, X. Boix, J. Hidary, & H. N. Mhaskar(2018) Theory of deep learning III: the non-overfitting puzzle. *CBMM memo 073*.

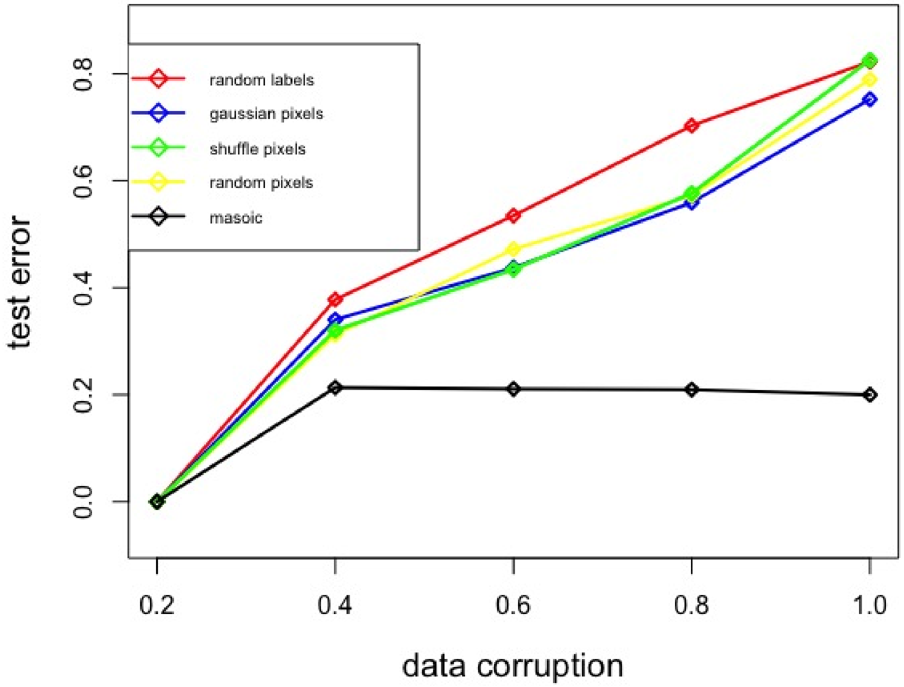

Figure 3: Test error under different data corruptions.

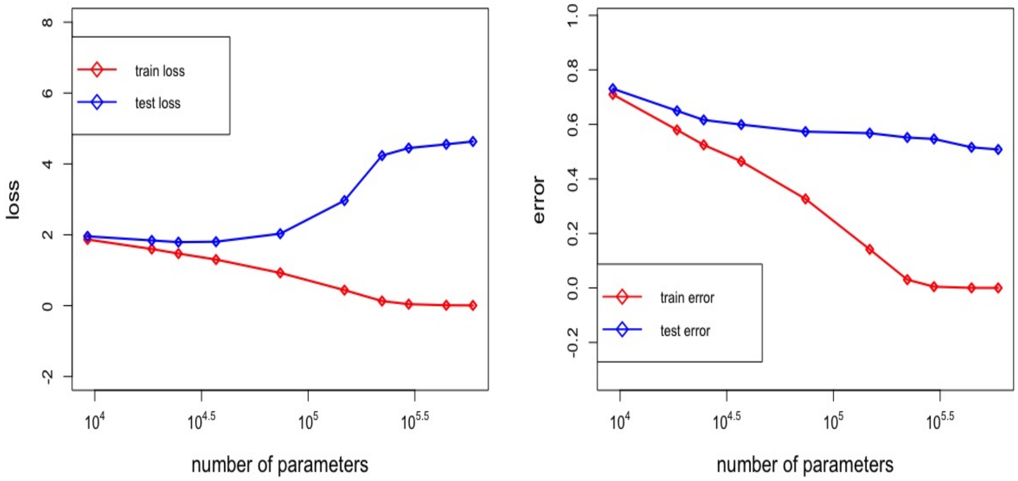

Figure 4: When training error becomes zero, test error (misclassication) does not increase (resistance to overfitting) but test loss increases showing overfitting as model complexity grows.

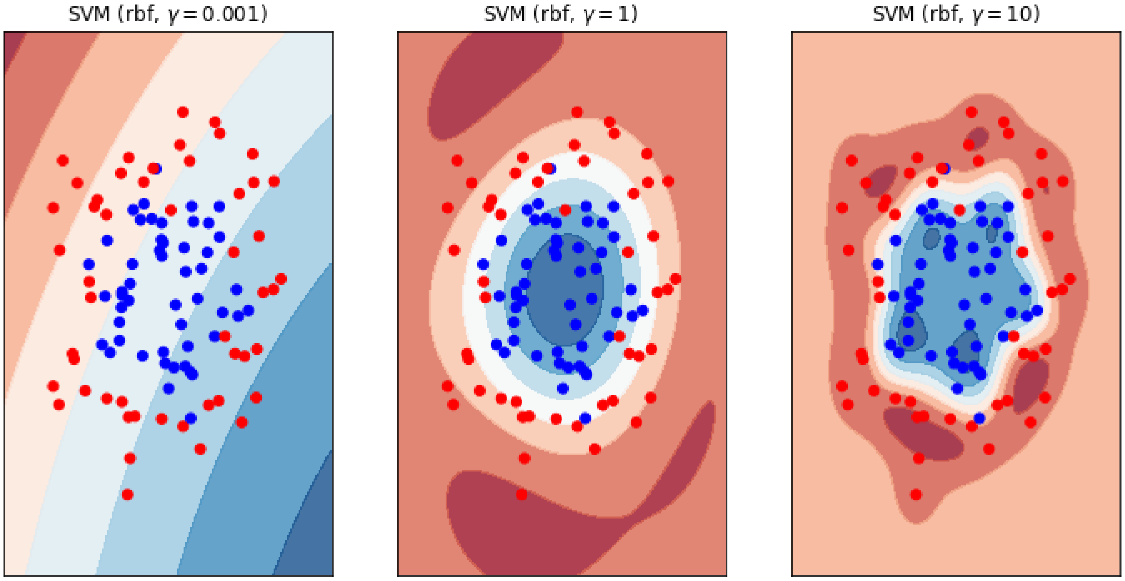

Figure 5: Left: Underfitting Right:Overfitting

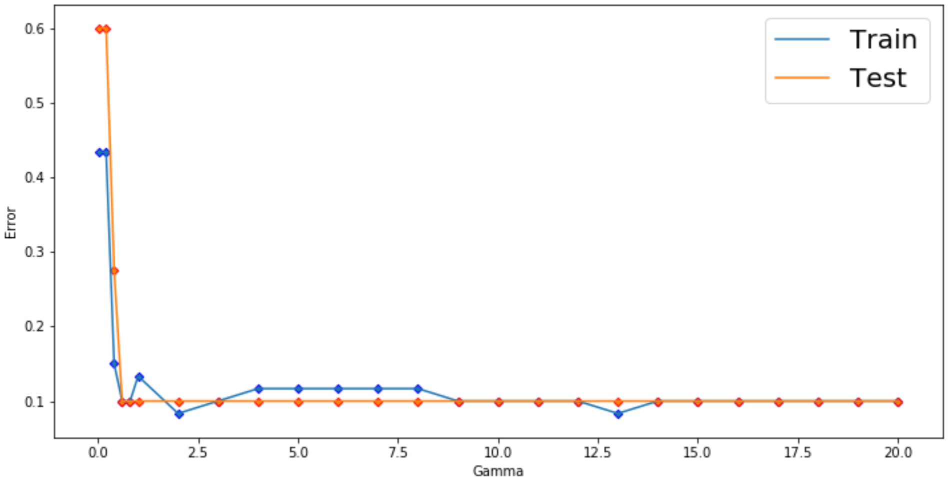

Figure 6: Nonfitting in classification error view