Generalizing without Overfitting: A Case Study on CIFAR-10 Dataset

Beijing Fang
Department of Civil and
Environmental Engineering
hippo@connect.ust.hk

Huangshi Tian
Department of Computer Science
and Engineering
htianaa@connect.ust.hk

Yunfei Yang
Department of Mathematics
yunfei.yang@connect.ust.hk

1 Introduction

The numbers of parameters in deep neural networks are often far more than the training samples. However, unlike traditional statistical learning, it seems that deep learning does not exhibit overfitting. To get a better understanding of this phenomena, we re-implement some experiments in [1, 2]. In the first experiment (section 2), we apply some modifications to the CIFAR-10 dataset and then train ResNet-18 to fit these datasets. In the second experiment (section 3), we train CNN on CIFAR-10 classification problem. We increase the number of parameters gradually and record the related loss and classification accuracy. Finally, we analyze and discuss our results in section 4.

2 Capacity of neural network

2.1 Experiment

To test the capacity of deep neural network, we use ResNet-18 to fit the standard and modified CIFAR-10 datasets. As in [2], we run our 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.
- **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.

2.2 Results

Figure 1 shows the learning curves of the ResNet-18 model on the CIFAR-10 dataset under various settings. Due to the limit of time, we only train each model 100 epochs. We observe that the training loss will always decay to zero (at least very close to zero. Although we do not achieve zero loss on some settings, the losses seem to decrease slowly). Since the loss of true labels decays to zero very fast, we only train the model 50 epochs in this case. We also tried to train the random labels (i.e. $p = 1$), but the loss decreases very slow so we do not plot the curve here.

Figure 2 and 3 are the test loss and test accuracy of corrupted labels, respectively. Note that the loss and accuracy are computed on the **true labels** test set. Figure 2 shows that for $p = 0.2$ and $p = 0.5$,

Preprint. Work in progress.

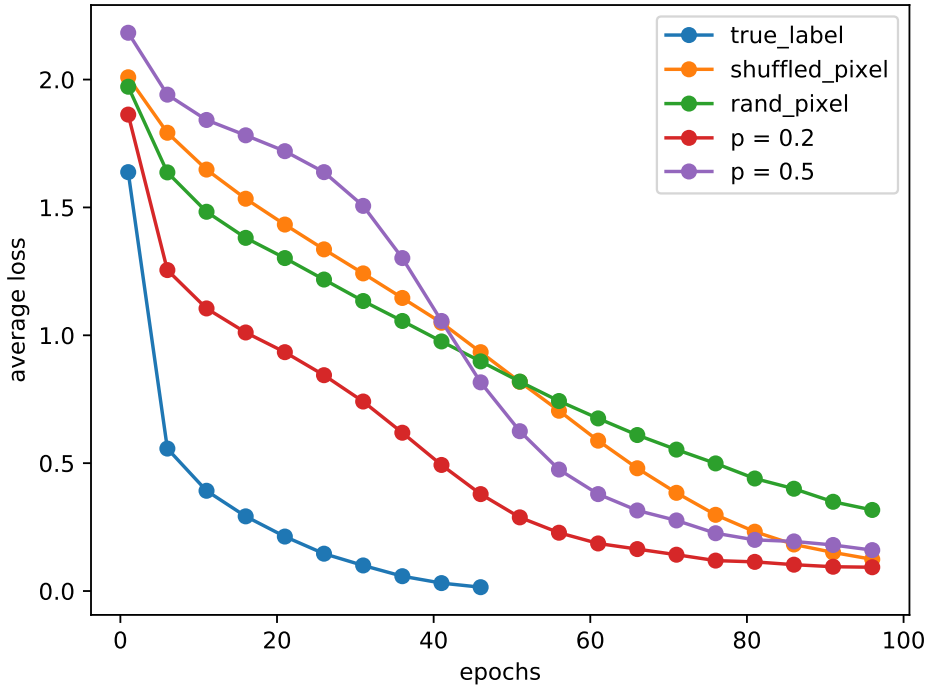

Figure 1: Fitting random labels and random pixels on CIFAR-10.

the model has exhibited overfitting. We observe that the test accuracy firstly increase, then decrease and finally become steady. We will provide some possible explanations for this phenomenon in section 4.

3 Non-overfitting puzzle

3.1 Model

We follow [1] and adopt an all-convolution architecture. Specifically, we first put together five convolutional layers, then connect the last one to a fully connected layer which has ten outputs, and take them as the model output. For the input channel of the first layer and the output channel of the last one, we fix them as three. The kernel is applied with stride 1 and without any padding.

To vary the number of parameters, we change the number of channels, n, and the size, s, of kernels used in convolution. To simplify the setting, we make all the internal channels among those convolutional layers have the same number. As each channel of each layer has its own kernel, the number of kernel parameters is $(3n^2 + 6n)s^2$. The fully connected layer has input of size $[32 - 5(s - 1)]^2$ because the image size is 32×32 and each convolutional layer will reduce the size by $s - 1$. $10 \cdot (37 - 5s)^2$. Henceforth, we control the number of parameters through n and s.

3.2 Experiment

Owing to overly long training time, we only use 1/10 of CIFAR-10 dataset, i.e., 5000 images, as training samples in our experiment. The learning rate is set to 0.01 and no exponential decay or other acceleration technique is used. For each model, we train it with 500 epochs and record the loss value and error rate. The error rate is obtained with tests on 1000 images, 1/10 of the original CIFAR-10 testset.

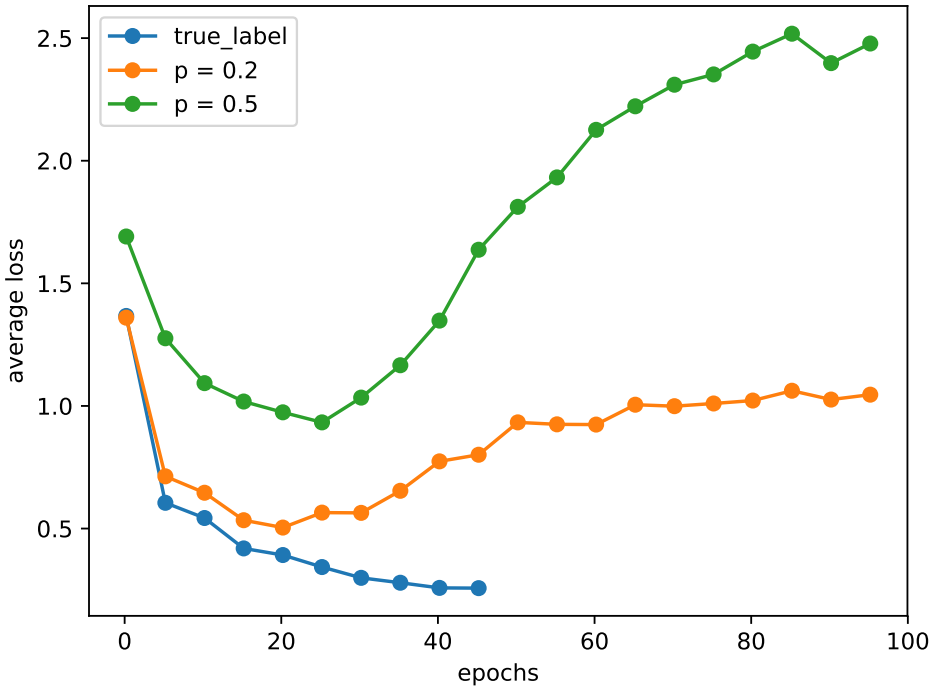

Figure 2: Test loss of corrupted labels.

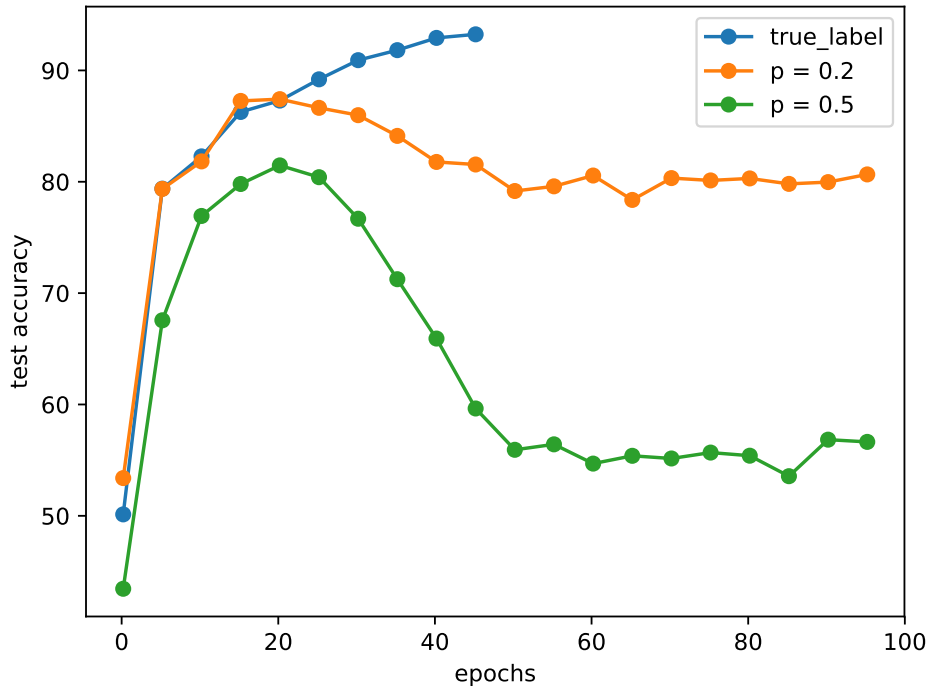

Figure 3: Test accuracy of corrupted labels.

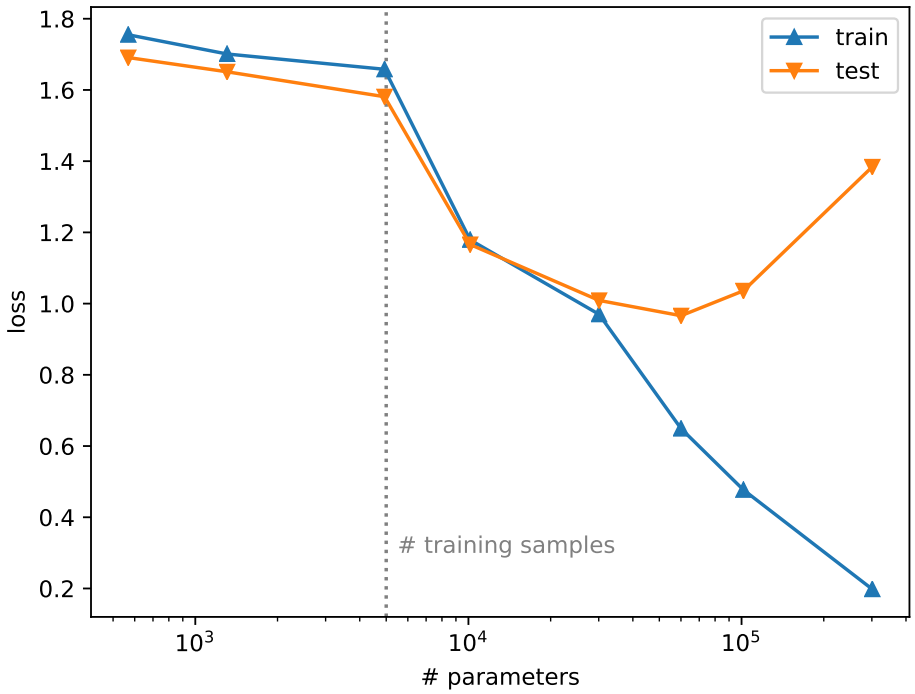

Figure 4: With an increasing number of parameters, the test loss appears to be overfitting.

3.3 Results

Figure 4 and 5 present the result of our experiments. As clearly shown, the loss value is severely overfitted. When the number of parameters exceeds that of training samples, the training loss quickly drops but the test loss grows in a reverse direction. This conforms with the traditional wisdom that over-parameterized models tend to overfit. However, the test accuracy doesn't show such trend. It keeps going up with the training accuracy, indicating that the model is still well better with overmany parameters.

4 Discussion

There are many interesting phenomena in the experiments and results, we describe them and give possible explanations as follow.

- The results of the first experiment show that deep neural networks can fit CIFAR-10 very well, even when we modify the dataset by random noises. The convergence time will increase steadily as we increase the noise level. Our explanation for this phenomenon is that CNNs have sufficient capacity to 'memorize' the entire training set. When a random transform is applied to the data, it is likely to make the landscape of the loss more complicated. So the optimization algorithm will need more time to converge, but, due to the capacity of CNNs, the global minimum is still zero.
- When doing the first experiment, we observe that the convergence rate highly depends on the optimization algorithms. But it seems that no matter which optimization methods we use, the loss of the original dataset decreases faster than modified datasets. It means that the original loss function has a simpler landscape. This observation also shows that a random transform on the data is more likely to complicate the landscape of the loss function.

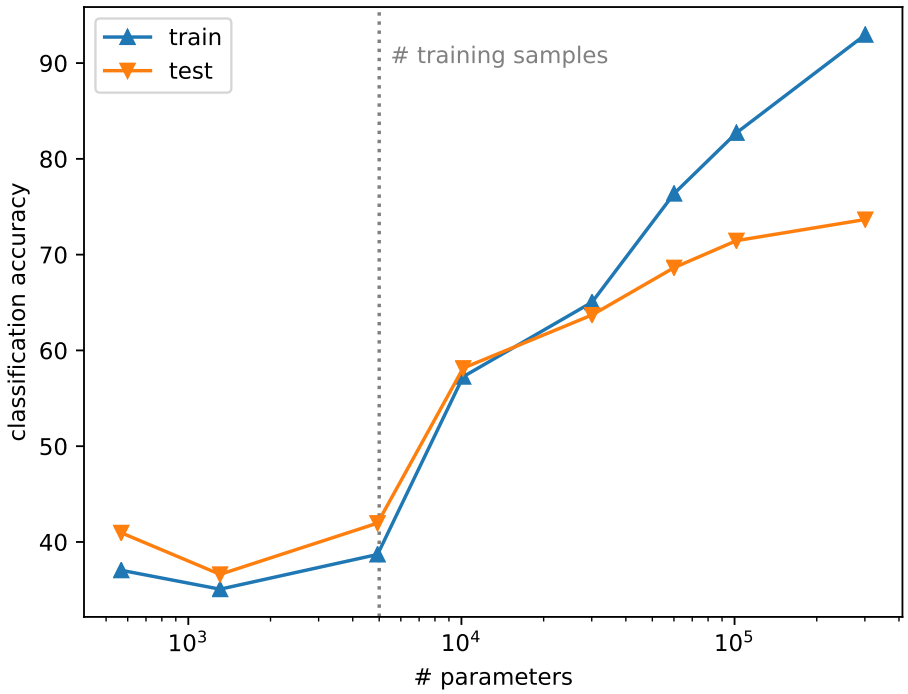

Figure 5: The test accuracy doesn't overfit as the number of parameters increases.

- In figure 3, we see that the accuracy firstly increase, then decrease and finally become steady. One possible explanation is that the models learn the 'true structure' of data first and then learn the random noise. So the test accuracy on the **true labels** will increase when the model is learning the 'true structure'. But, as the model learn the noise gradually, the accuracy decrease to the noise level.
- The second experiment shows that as the number of parameters increases, the model will gradually overfit the data, which conforms with traditional machine learning. But the question is why accuracy does not overfit? We think there are two possible reasons: The non-overfitting is a property of deep neural networks which we do not understand yet. Or it is caused by special structure of CIFAR-10. However, base on our experiments, we can not assert whether this phenomenon is caused by the dataset or the network. This could be an interesting topic for future research.

Contribution

In this project, we have jointly discussed the overall plan. Individually, Beijing Fang programs in the first experiment, Huangshi Tian does the second experiment and Yunfei Yang helps with experiments and analyzes the results.

References

[1] T. Poggio, K. Kawaguchi, Q. Liao, B. Miranda, L. Rosasco, X. Boix, J. Hidary, and H. Mhaskar. Theory of deep learning iii: the non-overfitting puzzle. Technical report, CBMM memo 073, 2018.

[2] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. *arXiv:1611.03530*, 2016.