MATH6380O Project 2

Behind the Non-Overfitting Phenomenon of CNNs

DENG, Yizhe, GU Hanlin, HUANG Yifei, SUN Jiaze Department of Mathematics, HKUST

Introduction

Convolutional Neural Networks (CNNs) have been widely adopted for image recognition owing to their enormous expressive capacity. Yet contrary to traditional statistical learning intuitions, CNNs do not seem to exhibit overfitting. In this project, our objectives are to answer the following questions:

A. Can CNNs 'learn' random data?

B. How does model complexity affect overfitting?

We shall offer our own analysis on the phenomenon.

Data

Our experiments were conducted on CIFAR-10, which consists of 60,000 RGB images of size 32x32x3, where 50,000 are for training and 10,000 are for testing.

Figure 1. A few examples from CIFAR-10

Methodology

Objective A:

There are 2 ways to produce random data:

I. Randomly shuffling the **labels** of the entire or part of the training set

II. Randomly shuffling the **pixels of each image** in the training set

For confidence, we performed this experiment with various well-known CNNs, including **AlexNet, VGG-16,** and **Resnet-18.**

Objective B:

We employed a basic CNN (rather than ResNet-18 and so forth) for this task, its architecture is as follows:

- **5 convolutional layers, 3x3 filters, stride 2;**
- **Batch normalization after each convolutional layer;**
- **ReLU activation.**

We control the model complexity by either **increasing or decreasing** the number of filters.

Objective A Results

In this experiment, we use the ordinary SGD, a learning rate of 0.1, a batch size of 128, and train for 60 epochs.

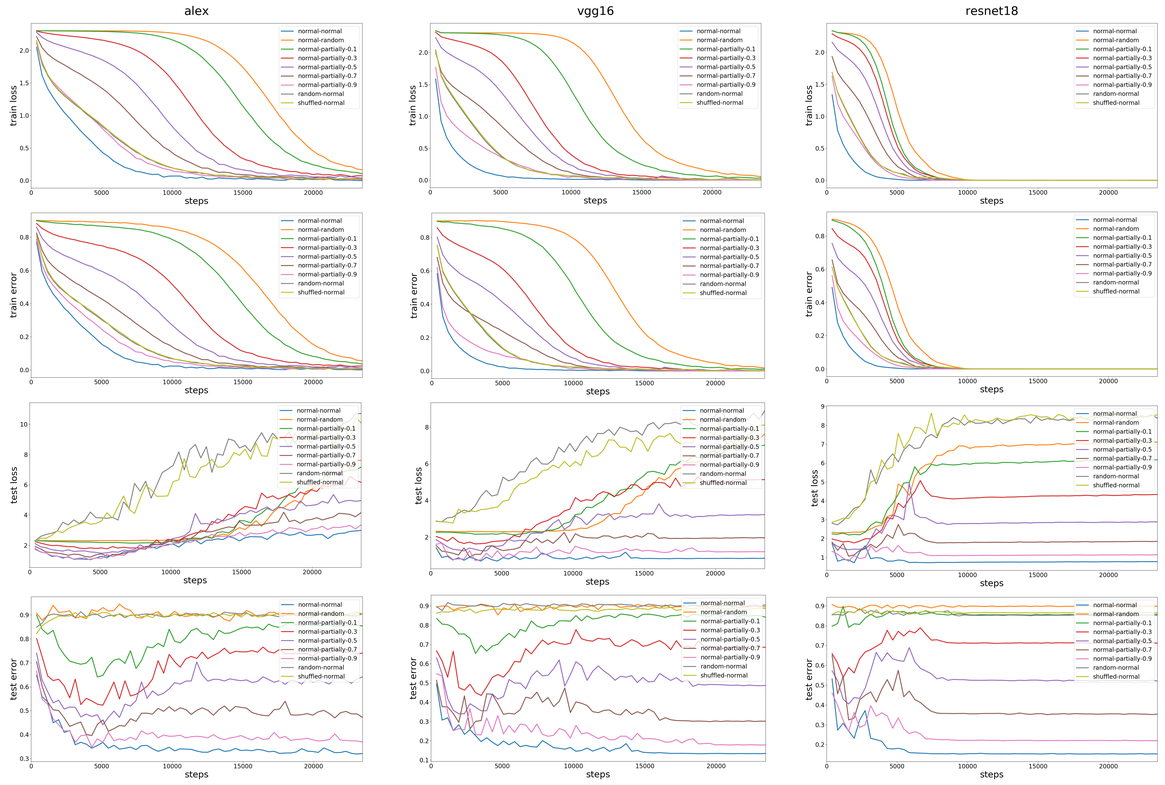

Figure 2. Results of Objective A. The legends above are in the '***t1*-*t2*-*keep***' format, where '***t1***' and '***t2***' refer to how we shuffle the training **images** and **labels** respectively. '***normal***' means no shuffling is applied; '***partially***' means some examples are not shuffled, and '***keep***' is the proportion of unshuffled examples. If ***t1*** = '***random***', then the pixels of each image are shuffled differently; if ***t1*** = '***shuffled***', then the pixels of each image are shuffled in the same way. If ***t2*** = '***random***', then all labels are shuffled randomly.

Objective B Results

In this experiment, we use the same settings as the ones in Objective A. We ran our basic CNN model 42 times, with a higher number of filters in each subsequent trial. In particular, the first trial has 3 filters and the last one has 350.

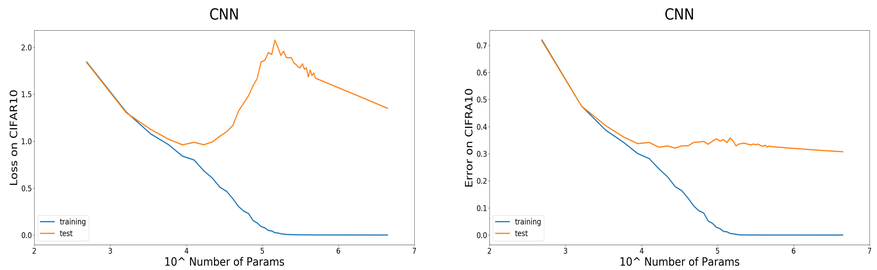

Figure 3. Results of Objective B.

Discussion

A. CNN 'memorizes' random data.

Phenomenon: CNNs can easily fit random data, with slight increases in training time depending on the shuffling method used.

Possible explanation: CNNs have sufficient capacity to memorize the entire training set. A random transformation on the data merely complicates the landscape of the loss function, thus increasing the training time but otherwise posing no significant hindrance to the eventual convergence.

B. Higher complexity leads to overfitting, but ONLY in loss

Phenomenon: As CNN complexity increases, test loss exhibits overfitting, while test accuracy remains steady.

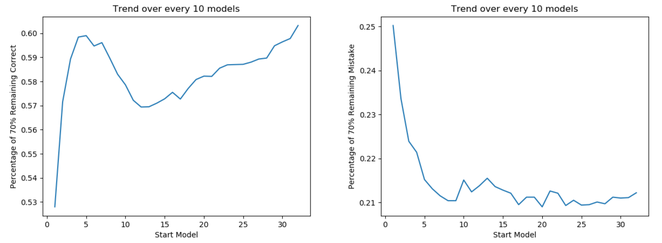

Figure 4. The proportion of examples that are classified correctly (left) or wrongly (right) 7 times out of 10 consecutive models of increasing complexity. In other words, we move a size 10 window over 42 models that are in ascending order of complexity, which is why the horizontal axes run from 0 to 33.

Possible explanation: It was expected that loss would exhibit overfitting, the question is why accuracy does not. Figure 4 shows the proportion of examples which are consistently classified correctly or incorrectly. From the graphs, one can see that around 60% of the test examples are consistently classified correctly, whereas 20% are consistently classified wrongly. A possible reason is that 60% of the test examples might be similar to the training ones, thus are easier for the model to give the right classifications. However, we still could not ascertain whether this phenomenon is caused by CIFAR-10 or the CNN itself, and this could be a potential direction for future research.

Acknowledgement

- **DENG Yizhe**: Analyzing the results
- **GU Hanlin**: Analyzing the results
- **HUANG Yifei**: Running the models and analyzing the results
- **SUN Jiaze**: Analyzing the results and writing the poster

References

- Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. *Understanding deep learning requires rethinking generalization*. Nov 10, 2016.
- Tomaso Poggio, K. Kawaguchi, Q. Liao, B. Miranda, L. Rosasco, X. Biox, J. Hidary, and H. Mhaskar. *Theory of Deep Learning III: the non-overfitting puzzle*. Jan 30, 2018.