AI6103 Homework Assignment

Li Boyang, Albert

1 Introduction

In this homework assignment, we will investigate the effects of hyperparameters such as initial learning rate, learning rate schedule, weight decay, and data augmentation on deep neural networks.

One of the most important issues in deep learning is optimization versus regularization. Optimization is controlled by the initial learning rate and the learning rate schedule. Regularization is controlled by, among other things, weight decay and data augmentation. As a result, the values of these hyperparameters are absolutely critical for the performance of deep neural networks.

For simplicity, in this assignment, we do not require you to divide the training set further into a training set and a validation set. However, you should know that the three-way split is necessary for proper and rigorous performance evaluation. You will be required to do that for the group project.

The report should be in the double-column AAAI format, for which the author kit can be downloaded from https://aaai.org/Publications/Author/socs-submit.php. The LaTeX format is preferred to the Word format, though the latter is allowed. The report should contain six pages or less, excluding references. Exceeding the page limit will automatically result in deduction of 20 points. Modifying the report format to avoid exceeding the page limit will similarly result in deduction of 20 points.

The following requirements apply to all experiments in this homework.

First, you should use the ResNet-18 network and the CIFAR-10 dataset. You should use the SGD optimization algorithm with momentum set to 0.9. You should not use other optimization algorithms like Adam or Adagrad. The ResNet code has been provided as a Colab notebook on NTULearn, which you may need to modify for this assignment.

Second, you should draw the following diagrams: (1) training loss and test loss against the number of epochs, and (2) training accuracy and test accuracy against the number of epochs. These diagrams allow us to analyze the training trajectory intuitively, which is critical in the diagnosis of deep neural networks. Example code for drawing these diagrams can be found in the code file for logistic regression on NTULearn.

Third, you are required to describe the empirical results with words. Even if the diagrams contain all the information, it may not be immediately clear what the most important findings are. You need to point them out to the reader. After that, you should discuss possible reasons for the empirical observations, possibly by relating them to materials discussed in the lectures.

2 Network and Dataset (10%)

Before training begins, a machine learning practitioner needs to develop a good sense of the model and the dataset. The more you know about them, the easier it is for you to debug and find solutions when things do not go as expected — they almost never go as expected in the first few trials. Describe the architecture of the ResNet-18 network and the CIFAR-10 dataset .

For the dataset, show some example images. You can investigate the data from multiple perspectives. Answer the following questions.

1. Are the object centered in the image or could they appear in the corners?

2. Are there images where the objects are occluded?

3. Are there other objects in the images?

Feel free to ask new questions and answer them yourself. This section accounts for 10% of the total score.

3 Learning Rate (20%)

We will first investigate the initial learning rate. Run three experiments with the learning rate set to 0.1, 0.01, and 0.001 respectively. The batch size should be set to 128. You should use neither weight decay nor learning rate schedule. For data augmentation, you should use random cropping and random horizontal flip as in the code on NTULearn. Train the networks for 15 epochs under each setting.

Report the final losses and accuracy values for both the training set and the test set. Plot the training curves as described in the introduction. Which learning rate performs the best in terms of training loss and training accuracy? Which learning rate performs the best in terms of test loss and test accuracy? Discuss possible reasons for the phenomena you observe. This section accounts for 20% of the total score.

4 Learning Rate Schedule (30%)

Next, we gradually decrease the learning rate. One effective learning rate schedule is cosine annealing. Describe this particular schedule intuitively and with one or more mathematical equations (10%). This part accounts for 10% of the total score.

When we adjust the learning rate, we look for one that minimizes training loss. Using this criterion, identify the best learning rate from the last experiment. Use this as the initial learning rate and keep all other settings and hyperparameters unchanged. Conduct experiments under two settings: (1) train for 300 epochs with the learning rate held constant, and (2) train for 300 epochs with cosine annealing, which decreases the initial learning rate to zero over the entirety of the training session.

Report the final losses and accuracy values for both the training set and the test set. Plot the learning curves and describe your findings. Discuss possible reasons for the differences in the two experimental conditions. This part accounts for 20% of the total score.

5 Weight Decay (20%)

Weight decay is similar to the L2 regularization used in Ridge Regression. For model parameter $w \in \mathbb{R}^n$ and an arbitrary loss function $\mathcal{L}(w)$, we add the regularization term $\frac{1}{2}\lambda\|w\|^2$ to the loss and optimize the new loss function $\mathcal{L}'(w)$

$$w^* = \arg\min_w \mathcal{L}'(w) = \arg\min_w \mathcal{L}(w) + \frac{1}{2}\lambda\|w\|^2. \tag{1}$$

Applying gradient descent on $\mathcal{L}'(w)$ leads to the following update rule,

$$\begin{aligned} w_{t+1} &= w_t - \eta\left(\frac{\partial\mathcal{L}(w_t)}{\partial w_t} + \lambda w_t\right) && (2)\\ &= w_t - \eta\frac{\partial\mathcal{L}(w_t)}{\partial w_t} - \eta\lambda w_t && (3) \end{aligned}$$

The above shows that, instead of gradient descent on $\mathcal{L}'(w)$, we can perform gradient descent on $\mathcal{L}(w)$ and subtract $\eta\lambda w$ from the current w in each update. Directly applying the subtraction on w is called weight decay. Surprisingly, weight decay often outperforms L2 regularization. For further reading (not required for this assignment), see [1].

Add weight decay to the best experimental setting discovered so far (you need to decide what is the best setting, which may include the best learning rate and the cosine schedule). Experiment with two different weight decay coefficients λ, 5×10^{-4} and 1×10^{-2}, and illustrate their regularization effects using training-curve diagrams. Report the final losses and accuracy values for both the training set and the test set. The network should be trained for 300 epochs. This section accounts for 20% of the overall score.

6 Data Augmentation (20%)

In Lecture 5, we examined a few data augmentation techniques such as random horizontal flip, random cropping, and cutout / random erasing.

With the best experimental setup you discoverd so far, experiment with the Cutout augmentation technique [2] using `torchvision.transforms.RandomErasing`. The `value` argument should be set to the respective means of the three color channels, calculated across the entire training set, so as to minimize the effects of the augmentation on the image distribution. Train the network for 300 epochs. Report the final losses and accuracy values for both the training set and the test set. Show the effects of this augmentation technique with diagrams and describe them in English. Discuss possible reasons for these effects. This section contributes to 20% of the overall score.

7 Grading Criteria

This assignment will be graded using the following criteria:

- You can perform the experiments correctly, as demonstrated in the results.
- You can plot the experimental results correctly and in an easy-to-understand manner.
- You can describe the results of the experiments accurately and concisely.
- You can analyze and explain the results, and correctly relate the results to content discussed in the lectures. Note that enumerating everything in the lectures indiscriminately will result in point deduction.
- You can write a report that demonstrates correct usage of English. You can communicate clearly, concisely, and unambiguously within the page limit. Remember, it takes more effort to write a short report that conveys all the important points than a long report.

References

[1] G. Zhang, C. Wang, B. Xu, and R. Grosse, "Three mechanisms of weight decay regularization," in *ICLR*, 2019.

[2] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv 1708.04552*, 2017.