AI6103 Homework Assignment (v1.2)

Li Boyang, Albert

September 13, 2021

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 all experiments in this homework, 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.

In all the experiments, you will 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 code has been provided as a Colab notebook on NTULearn, which you can modify as you see fit.

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.

2 Network and Dataset

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 (10%).

For the dataset, show some example images. You can investigate the data from multiple perspectives. Here are some example questions: are the object centered in the image or could they appear in the corners? Are the objects occluded? Are there other objects in the images? Feel free to ask new questions and answer them yourself.

3 Learning Rate

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. 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 (20%).

4 Learning Rate Schedule

Next, we let the learning rate gradually decrease. One effective learning rate schedule is cosine annealing. Describe this particular schedule mathematically and intuitively (10%).

When we adjust learning rate, we look for one that minimizes training loss. Using this criterion, identify the best learning rate the last experiment. Use this as the initial learning rate and keep other 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.

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. (20%)

5 Weight Decay

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,

$$w_{t+1} = w_t - \eta\left(\frac{\partial\mathcal{L}(w_t)}{\partial w_t} + \lambda w_t\right) \tag{2}$$

$$= w_t - \eta\frac{\partial\mathcal{L}(w_t)}{\partial w_t} - \eta\lambda w_t \tag{3}$$

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 experimental settings used previously (including 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 (20%).

6 Data Augmentation

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

Use the best experimental setup you discoverd so far (including the best learning rate, the cosine schedule, and weight decay). Implement the Cutout augmentation technique [2] using `torchvision.transforms.RandomErasing`. The `value` argument should be set to the means of the three color channels, calculated across the entire dataset, so as to minimize the effects of the augmentation on the image means. 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 (20%).

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.
- Importantly, 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 some point deduction.
- You can write a report that demonstrates correct usage of English.

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.