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.

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 result in deduction of 20 points.

The following requirements apply to all experiments in this homework.

First, you should use the MobileNetV2 network and the CIFAR-100 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 MobileNetV2 code has been provided on NTULearn, which you need to modify for this assignment.

Second, you should draw the following diagrams: (1) training loss and validation loss against the number of epochs, and (2) training accuracy and validation 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 Data Preprocessing (10%)

Before training begins, a machine learning practitioner needs to develop a good sense of the data (and the network). 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.

The CIFAR-100 training set you get from `torchvision` contains 50,000 images. Randomly divide this dataset into a new training set and a validation set, containing 40,000 and 10,000 data points respectively. Use random seed 0 for the partitioning. Show the following in your report.

1. The line(s) of code you use to partition the data.

2. The proportion of each class in the new training set.

Compute the mean and standard deviation for each color channel on the training set. Report these numbers and use them in your preprocessing pipeline to whiten / normalize the data. You should also use standard random horizontal flip and random cropping as data augmentation. The horizontal flip probability is 0.5. For random cropping, first apply 4-pixel paddings on all sides and crop a 32-by-32 patch from the image. Show the code in your report.

In Sections 3-6, use performance on the validation set to tune the hyperparameters. You should only look at the test set performance at the end of all experiments.

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 0.5, 0.05, and 0.01 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 described in the previous section. Train the networks for 15 epochs under each setting.

Report the final losses and accuracy values for both the training set and the validation 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 validation loss and validation accuracy? Identify the best learning rate that minimizes the training loss. Discuss possible reasons for the phenomena you observe.

This section accounts for 20% of the total score.

4 Learning Rate Schedule (20%)

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 (5%).

Use the best learning rate identified earlier 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 validation set. Plot the learning curves and describe your findings. Discuss possible reasons for the differences in the two experimental conditions. This part accounts for 15% 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 learning rate you discovered, and the cosine learning rate schedule. Other configurations should remain identical to the previous experiment. Experiment with two different weight decay coefficients $\lambda = 5 \times 10^{-4}$ and 1×10^{-4}, and illustrate their regularization effects using training-curve diagrams. Report the final losses and accuracy values for both the training set and the validation set. The network should be trained for 300 epochs. This section accounts for 20% of the overall score.

6 Data Augmentation (30%)

The lectures introduce a few data augmentation techniques such as random horizontal flip, random cropping, and mixup.

With the best experimental setup you discovered so far, experiment with the mixup augmentation technique [2]. Set the hyperparameter α to 0.2. Draw the probability density function associated with the beta distribution parameterized by this α (5%).

Train the network for 300 epochs. Report the final losses and accuracy values for both the training set and the validation set. Show the effects of this augmentation technique with diagrams and describe them in English. Discuss possible reasons for these effects. (20%)

Finally, report the accuracy on the hold-out test set (5%). This is the accuracy that you should expect the trained model to perform at for all similar images in the future.

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 *The 7th International Conference on Learning Representations*, 2019.

[2] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *The 6th International Conference on Learning Representations*, 2018.