AI6103 Homework Assignment

Li Boyang, Albert

0 Policy on Generative AI

You are allowed to use generative AI and large language models, such as ChatGPT, to help you with the project. However, before you begin, you should be aware of known issues with these technologies. They tend to make up false statements that sound convincing (i.e., hallucinate), make mathematical errors, give the same or very similar answers under similar prompts (cf., [1]), and generate verbose text full of cliche but little information. Ultimately, it is you who are responsible for your own report. "GPT wrote it" will not be accepted as a valid excuse if your report contains substantial textual overlap with other students' reports, contains factual and mathematical errors, or gives incomplete answers to the questions.

1 Introduction

In this homework assignment, we will investigate the effects of hyperparameters such as initial learning rate, learning rate schedule, weight decay, and activation function 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 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/authorkit24-2/. 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 MobileNet 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 MobileNet 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 verbally. 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.

Lastly, you should use performance on the validation set to tune the hyperparameters. You should only look at the test set performance at the end of Section 5.

2 Provided Code Files

We provide two Python files, `model.py` and `data.py`. In `model.py`, we provide a modified MobileNet that may use ReLU or sigmoid as the activation function. You should use the ReLU version for most sections except Section 6. In `data.py`, we provide functions for creating the training set, the validation set, and the test set, and the associated data augmentation techniques.

3 Learning Rate (20%)

We will first investigate the initial learning rate. Run three experiments with the learning rate set to 0.2, 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 (25%)

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 20% of the total score.

5 Weight Decay (25%)

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 [2].

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.

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.

6 Activation Function (20%)

Change the activation function used in network blocks 4-10 to the sigmoid activation function. This can be achieved by setting `sigmoid_block_ind` to `[4,5,6,7,8,9,10]` in the constructor of `MobileNet`. Train the network using the best hyperparameters you discovered for 300 epochs. Plot the training and validation diagrams and compare that with the ReLU version.

Further, plot the 2-norm of the gradient vector of `model.layers[8].conv1.weight` over the 300 epochs. Describe the observations and explain them by relating to the class content. In order to discourage copying from reports of previous semesters, if your report does not contain this section, the instructor may deduct 50% from your score. The point deductions will be at the complete discretion of the instructor.

7 The Takeaway Message (10%)

Please reflect on this assignment and describe what skills or tricks you learned and how you may be able to apply them in your own work in the future. The minimum length for this section is 50 words.

8 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] S. Jentzsch and K. Kersting, "ChatGPT is fun, but it is not funny! humor is still challenging large language models," *arXiv Preprint 2306.04563*, 2023.

[2] G. Zhang, C. Wang, B. Xu, and R. Grosse, "Three mechanisms of weight decay regularization," in *The 7th International Conference on Learning Representations*, 2019.