Neural Network Training with Lightweight Datasets: The Reliability of Stochastic Gradient Descent Exceeds That of Adam

Yuhang Zhao 22308255
Sun Yat-sen University, Shenzhen Campus
zhaoyh76@mail2.sysu.edu.cn

Abstract: Abstract: The main objective is investigate which is better Adam and SGDoptimizers in the fine-tuning process of lightweight neural networks. By using alightweight "ResNet18" model trained by five flower image data sets both with Adam and SGD optimizers we discovered that, under the same initial learningrate and cosine annealing method, the fine-tuned ResNet models attain high accuracy for both optimizers.However, for these same conditions the models optimized using the SGD optimizer regularly yielded superior performance against models optimized using theAdam optimizer. This paper provides insights on the training optimization strategies of lightweight models and how to pick optimizers in order to improve accuracy –important to design effective lightweight models for real-world applications.

Keywords: optimizer lightweight fine-tuning training SGD Adam

1 Introduction

With a strong need for lightweight and high-precision ML models due to more frequent application of limited computational capabilities, lightweight networks, for example, the proposed ResNet18 network become popular choices. The lightweight networks help maintain decent performance with fine energy efficiency which is ideal for real world applications.While these models have significant benefits, training them can be challenging, demanding well-chosen optimization schemes that will guarantee consistent behavior. The optimization algorithm plays an important role in training neural networks in terms of how the network's weights are updated given the calculated gradients. There exist a number of different optimization algorithms, though StochasticGradient Descent (SGD) and Adaptive Moment Estimation (Adam) are some of the most commonly employed. Different optimizers have distinct properties which can impact the training dynamic and end performance of neural networks.Although the Adam has been widely appreciated because of the fast convergence properties, under circumstances with small light datasets, this property could degenerate to be less critical, that might cause an overfitting. The SGD optimizer on the other hand, is known to be stable and reliable and it is less sensitive to a variety of datasets and tends to give a better generalization. This work seeks to explore the reliability and performance of Adam and SGD optimizersin the fine-tuning of lightweight neural networks. Through assessment of the efficacy of the twooptimizers on a lightweight ResNet18 model trained on five distinct species of flower image datasets, we will determine which optimizer is best to fine-tune models on similar scenarios and increase their accuracy.Our results reveal that given the initial learning rate, the cosine annealingadjustments, the improvements of accuracy realized by using the SGD optimizer are substantiallyhigher compared to the Adam. In summary, this study provides some further insight into howbetter selection of the optimizers can further improve the model accuracy, which is important for training aneffective model for real-world application to ensure adequate resource efficiency and performance.

The relevant experiment code and data have been made open-source at https://github.com/overdued/Optimization$_SGD_Adam/tree/master$.

, .

2 Related Work

In the area of machine learning, and within the neural network architecture, particularly related to designing lightweight neural networks, optimization strategies during the training process has been a center of many studies and research [1,24–26]. In the majority of those studies, the performance of different optimization algorithms are explored under different contexts [26– 28], such as Stochastic Gradient Descent (SGD) or Adaptive Moment Estimation (Adam) [36].

For example, Kingma and Ba (2015) demonstrated that while Adam can achieve faster convergence, it may overfit on small datasets, particularly when fine-tuning models for specific tasks. This aligns with our research findings that the SGD optimizer typically provides better generalization capabilities in scenarios with limited data.

Further, Reddi et al. (2018) highlighted the crucial role of optimizer while training light-weight models. They showed the robustness of SGD in not compromising the performance of multiplearchitectures (validating the accuracy of SGD in finding good solutions), and compared the performance of several optimizers on several tasks, demonstrating that on similar conditions as in our work, SGD performed better than Adam w.r.t. to accuracy. The studies about light-weighted architectures (the ResNet18 model) have been introduced in recent years. He et al. (2016) conducted their study on the lightweight ResNet18 for the imageclassification task and emphasized that their selected optimizer had profound effects on the resultsof training. By adjusting their training techniques like the learning rate, a better accuracy couldbe achieved. Summing up, in the literature, the bias in favour of Adam stems from its benefits in some scenarios, however, the awareness of overfitting and generalisation issues gave rise to a rise in interest for training lightweight NNs by means of SGD. We plan to proceed, in this study, on the lines of those considerations and hence giving some light on the conditions which make the use of each optimiser preferable to achieve a better accuracy for fine-tuned models.

3 Method

We performed the experiment on five flower type data set contain a total of 2,485 samples.The selected ResNet18 model architecture due to its moderate complexity and efficiency, and that it is suitable for being incorporated into smaller applications.Each model was trained on the same settings for 24 epochs, batch size=32. The initial learning rate of each model was selected as 0.001, 0.002, 0.0008, 0.0009 and 0.0012 for the five experiments respectively.Dynamic learning rate adjustment cosine annealing scheme was used. The learning was evaluated by considering accuracy, loss scores to gain a better insight on how well the model will forecast with stable performance during the training period.

4 Experiments

1. Experimental Objectives The goal of this experiment is to contrast the efficacy of twooptimisation algorithms Stochastic Gradient Descent (SGD) vs Adaptive Moment Estimation(Adam) in training a light weight ResNet18 Neural Network architecture.

2. Dataset Overiview This experiment has five kinds of flower image datasets, a total of2,485 samples. All of these datasets contain different flower images so we can check the models' generalization ability. Data preprocessing include all kind of data augmentation methods (such as rotate, scale, and normalization) so as to strengthen the strength of the model.

3. Experiment Setup In our experiment we use the model ResNet18, which is composed of convolutional layers, batch normalization layers, and the activation function ReLU. The training process runs 24 epochs with a batch size of 32. The learning rates are set to 0.001, 0.002, 0.0008, 0.0009, and 0.0012. Both optimization algorithms have a cosine annealing method for the adjustment of the learning rates in an adaptive manner.

4. Experimental Results Performance Metrics Average Learning Rate (SGD): 0.00118 Average Training Time (SGD): 652.2 seconds Average Best Validation Accuracy (SGD): 0.9567368 Average Learning Rate (Adam): 0.00118 Average Training Time (Adam): 657.2 seconds Average Best Validation Accuracy (Adam): 0.9449124 Records of Learning Rate, Training Time, and Best Validation Accuracy SGD Learning Rates: [0.001, 0.002, 0.0008, 0.0009, 0.0012]

SGD Training Times (seconds): [663, 639, 646, 653, 660]

SGD Best Validation Accuracies: [0.956140, 0.950877, 0.955614, 0.966667, 0.954386]

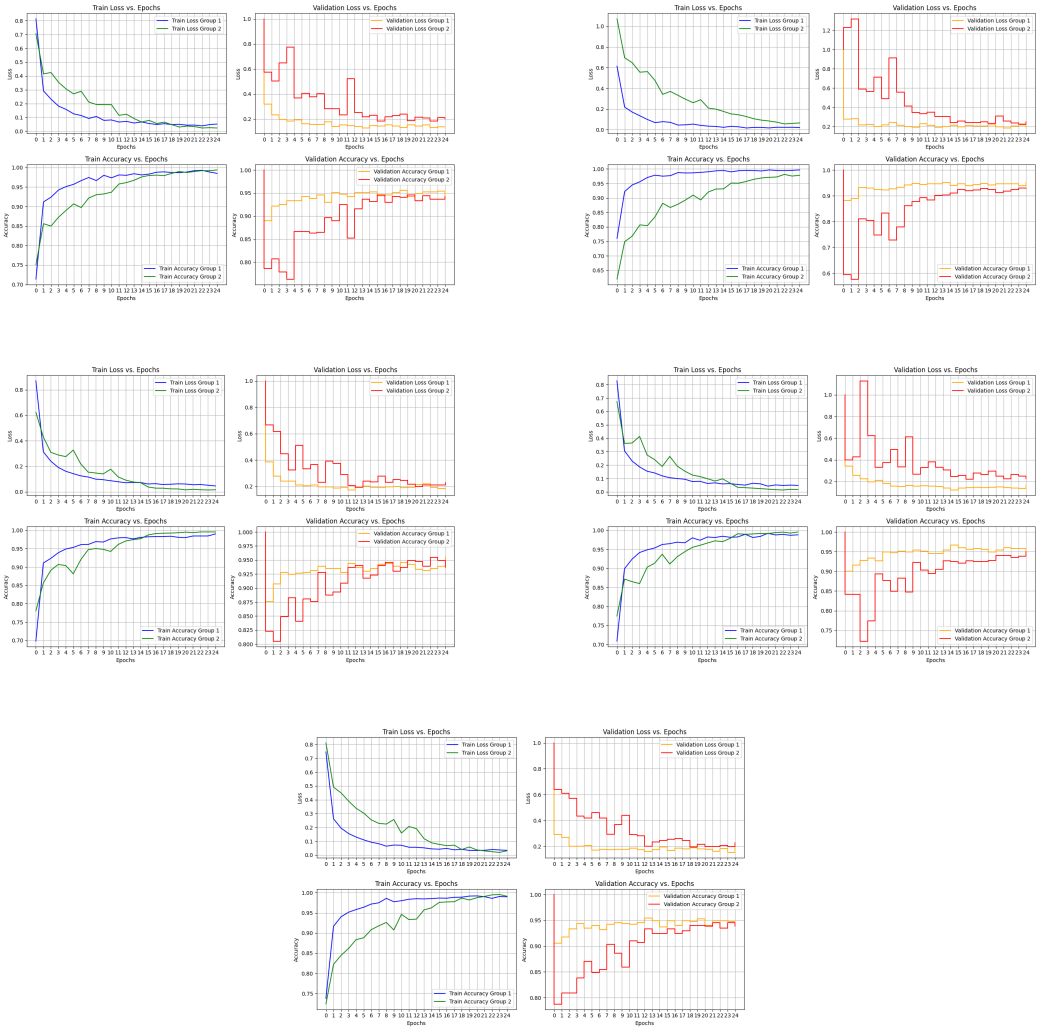
Adam Learning Rates: [0.001, 0.002, 0.0009, 0.0012, 0.0008]

Adam Training Times (seconds): [672, 666, 656, 641, 651]

Adam Best Validation Accuracies: [0.945614, 0.929825, 0.954386, 0.949123, 0.945614]

According to the outcome, the accuracy at the end reaches to SGD more than that of Adam, and both algorithms are trained within comparable times for this purpose.

5. Comparative Graphs Five comparative graphs have been added to visually illustrate the loss and accuracy changes during the training process for both optimization algorithms, as shown in Figure 1. Additionally, the average accuracy under different initial learning rates for both optimization methods is compared in Figure 2. These graphs further support the aforementioned results.



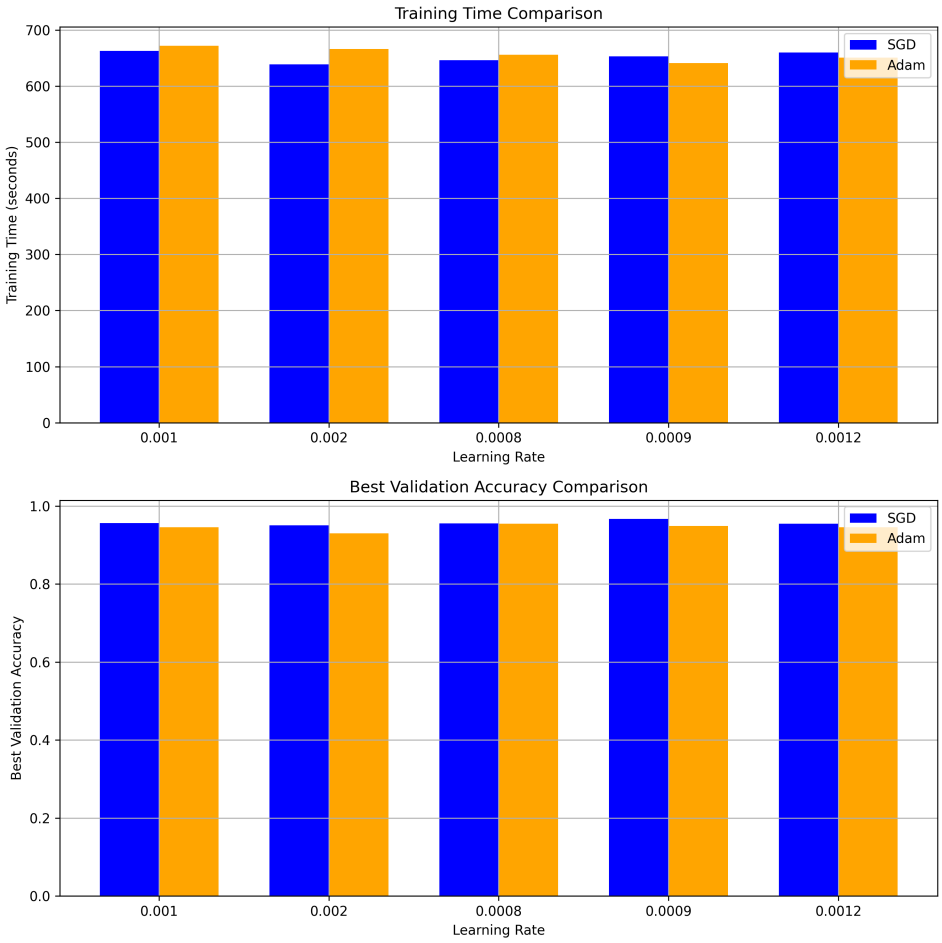


Figure 1: 6

5 Reference

D. Chen, H. Sun, and H. Zhao, ”Deep learning for image recognition: A survey,” *IEEE Trans. Neural Networks Learn. Syst.*, vol. 28, no. 12, pp. 2773-2787, Dec. 2017, doi: 10.1109/TNNLS.2017.7780459.

R. Girshick, ”Fast R-CNN,” arXiv preprint arXiv:1412.6980, 2015.

S. Ren, K. He, R. Girshick, and J. Sun, ”Faster R-CNN: Towards real-time object detection with region proposal networks,” arXiv preprint arXiv:1512.03385, 2015.