MODIFIED 9-LAYER RESNET WITH MIXED-PRECISION TRAINING FOR ACCELERATING TRAINING PHASE

Anonymous authors
Paper under double-blind review

1 INTRODUCTION

Deep Neural Networks (DNNs), as one of the leading machine learning methods, has achieved promising progress in many complex tasks, including image or speech recognition, gaming, and real-time language responses. Training of large DNNs, however, is still considered a time-consuming and computationally intensive task that demands extra high computational resources taking many days. The training phase involves a multitude of matrix-to-matrix additions and multiplications, which can be processed using large-scale parallelization available from Graphics processing units (GPUs). Therefore, it is our interest to reduce the complexity of the training process.

2 METHOD

2.1 SUPER-CONVERGENCE

According to Smith & Topin (2019), when using very large learning rates with the cyclical learning rate (CLR) policy, it shows a "super-convergence" phenomenon that can speed up training by as much as an order of magnitude. Therefore, we made a toy sample and tried hyperparameters as the paper suggest to train 18-layer ResNet. This model could achieve 85% validation accuracy on CIFAR-10 in 300 seconds without other data augmentations or data pre-processing techniques. Therefore, we argue that for this competition, adopting a super-convergence learning rate policy is essential.

2.2 MIXED-PRECISION TRAINING

Micikevicius et al. (2017) suggest that mixed-precision training has the potential to reduce the significant computational cost by performing operations in half-precision(16-bit floating-point), while storing minimal information in single-precision(32-bit floating-point) to retain as much information as possible in critical parts of the network. The Nvidia V100 general-purpose cores and tensor cores are designed for higher throughput with half-precision floating-point calculations, so reducing model precision greatly improves training speed. Therefore, we also applied the mixed-precision algorithm for our training to reduce floating-point computation and memory access time.

2.3 FINE-TUNED TOP-PERFORMING WORK IN DAWNBENCH COMPETITION

We noticed that top-performing training methods in DAWNBench competition (Coleman et al., 2019) can train a neural network 94% accuracy on CIFAR-10 in around ten seconds. Most of their models are based on 9-layer ResNet, which originate from Page (2018). The top 3 works all applied multiple GPUs to train the network which is not our interest, therefore, we referred to the work proposed by Page (2018) which also trained on a single Nvidia V100 GPU. Our work is a modified version based on its code. The most notable techniques from their work are the GPU-based preprocessing and batching of training data, 9-layer ResNet architecture, efficient BatchNorm implementation, earlier placement of max pooling, CeLU activation functions, and mirrored test time augmentation. The following are the major changes we made: (1) Modify structure and training and testing function to make it compilable for all datasets (not just CIFAR-10). (2) Increasing the training epoch to make full use of ten minutes of training time. (3) Because evaluation consists of classifying 10 classes only with 5000 training samples and 1000 testing samples, we sampled training samples in CIFAR-10 and only train in 5000 sampling samples instead of training in 50000 samples. We

tuned hyperparameters like batch size and learning rate policy using the reduced CIFAR-10 dataset and achieved 87% validation accuracy.
We also tested on the full CIFAR-10 and CIFAR-100 datasets, and this model could achieve 95% and 78% validation accuracy on CIFAR-10 and CIFAR-100, respectively in ten minutes.

3 Implementation Code

Code for our team submission is available at `https://github.com/AlexHoffman9/HAET-2021-resnet9`

References

Cody Coleman, Daniel Kang, Deepak Narayanan, Luigi Nardi, Tian Zhao, Jian Zhang, Peter Bailis, Kunle Olukotun, Chris Ré, and Matei Zaharia. Analysis of dawnbench, a time-to-accuracy machine learning performance benchmark. *ACM SIGOPS Operating Systems Review*, 53(1):14–25, 2019.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. Mixed precision training. *arXiv preprint arXiv:1710.03740*, 2017.

David Page. How to train your resnet, 2018. URL `https://myrtle.ai/learn/how-to-train-your-resnet-4-architecture/`.

Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, volume 11006, pp. 1100612. International Society for Optics and Photonics, 2019.