A Review on SimiGrad: Fine-Grained Adaptive Batching for Large Scale Training using Gradient Similarity Measurement

Arup Mazumder
Computer Science Dept.
Bauhaus University
Weimar, Thuringia, Germany
arup.mazumder@uni-weimar.de
arupseu@gmail.com

Disclaimer: I am not the writer of the original paper. I only reviewed the paper as part of the demonstration of my research interest and capability. Also, I want to mention that I am not an expert; instead, I want to consider myself a novice researcher in this field. Therefore, I assume that this review or part of this review will not be cited as a reference in future studies. The original paper is cited in this review. Please check the reference section.

***Abstract*—Large scale training is one of the challenging tasks in today's deep learning model that require a distributed training method to reduce the training time. But the distributed training method requires a large batch size that comes with the cost of generalization performance. Therefore, this paper came up with a noble technique called SimiGrad. SimiGrad is a lightweight automatic tool that follows an adaptive batching method at the mini-batch level to achieve state-of-the-art performance. Their proposed approach achieved promising results when using substantial batch sizes. In addition, they menioned that, SimiGrad can obtain a SQuAD score of 90.69 in BERT-Large pretraining with a batch size of 78k.**

I. STRENGTHS

Although I encountered some new concepts here, I enjoyed reading this paper because it was easy to follow. In addition, it explained the core concepts elaborately, which helped me dive deep. The related works section is also rich, which helped to distinguish how this paper is outstanding from others. Some of the key strengths are as follows:

A. Use of Cosine similarity metric

SimiGrad used the cosine similarity metric of the gradient of a mini-batch to measure the gradient variance[1]. In contrast, the previous studies are expensive in estimating gradients from different batch sizes because of the calculation of multiple batches as well as the process of collecting gradients from different batch sizes [2].

B. Learning Rate Scheduling

SimiGrad automatically adjusts the learning rate using a squire root factor of the batch size change where a warmup is not necessary. So SimiGrad can produce good performance even initial learning rate is selected randomly [1].

C. Phenomenal empirical study

The hardware used for the SimiGrad experiment is Ubuntu 18.04 on NVIDIA DGX-2 nodes (16 V100 GPUs). SimiGrad performance compared with state-of-the-art adaptive batching for performance evaluation, where SimiGrad achieves a record-breaking large batch size of 78k for BERT-Large pretraining with a SQuAD score of 90.69 [1].

II. WEAKNESS

A. Machine learning frameworks

In section 4 [1], the "mainstream machine learning frameworks" term is used. But why the name of those frameworks are not mentioned?

B. Training time

In section 5.4 [1], what is the dependency of the training time? Also, is the training time influenced a lot if we increase or decrease GPU cluster size and GPU-GPU bandwidth within the cluster?

REFERENCES

[1] Heyang Qin, Samyam Rajbhandari, Olatunji Ruwase, Feng Yan, Lei Yang, and Yuxiong He. Simigrad: Fine-grained adaptive batching for large scale training using gradient similarity measurement. In *NeurIPS 2021*, November 2021.

[2] Sam McCandlish, Jared Kaplan, Dario Amodei, and OpenAI Team. An empirical model of large-batch training, 12 2018.