Increasing Mini-Batch Size While Preserving Accuracy for Distributed Deep Learning

MARIE MCCORD, Middle Tennessee State University, USA
SAJAL DASH, NCCS, Oak Ridge National Laboratory, USA
FEIYI WANG, NCCS, Oak Ridge National Laboratory, USA
ARJUN SHANKAR, NCCS, Oak Ridge National Laboratory, USA

Increasing complexity of deep-learning models and scale of training datasets has yielded impressive and rewarding results, but these advances have also increased training times. An effective way to reduce training time is to use data-parallel distributed training across multiple workers. Each worker in distributed training requires enough work to justify the increased communication overhead and establish a balanced computation-to-communication ratio. This issue can be addressed by using large mini-batch sizes, which refer to the total chunk of training data that is processed through the network during a single iteration. Large mini-batch sizes provide each worker with a significant amount of work and reduce communication overhead by reducing the number of training iterations. However, beyond a certain size limit, the benefit of large mini-batch sizes comes at the cost of accuracy. To understand the effects of large mini-batch sizes on accuracy, we identified two techniques from prior research that scaled mini-batch size to some extent: *gradual warmup* and *linear rate scaling*.

Gradual warmup initializes training with a low and safe learning rate that is steadily increased to the desired rate over the first few epochs. Linear rate scaling is a technique that scales the learning rate by k when the mini-batch is scaled by k. We analyzed both techniques by experimenting with the ResNet-50 model, which was trained for the standard 90 epochs on the ImageNet dataset.

To analyze gradual warmup, we ran a series of experiments for mini-batch sizes 7,680 and 30,720 with 0, 5, 10, and 20 warmup epochs along with linear rate scaling. Training without a warmup produced the worst accuracy rates. For mini-batch size 7K, the optimal number of warmup epochs is 5. For mini-batch size 30K, the optimal number of warmup epochs is 10. Accuracy decreases past the optimal number of warmup epochs. Early training phases are generally sensitive to high learning rates due to extreme gradient changes from random weight initialization. A gradual warmup with an initialized low learning rate that takes small steps is necessary to stabilize the network during early training. However, once the gradient changes become less extreme after the first few epochs, low learning rates impede training by updating the weights too slowly.

Since we used linear rate scaling, which scales the learning rate by k when mini-batch size is scales by k, it seems reasonable that the optimal number of warmup epochs is linearly scaled as well. A larger learning rate would require more epochs to increase from the initialized low learning rate to the desired rate. So, a mini-batch size of 7K needs fewer warmup epochs than a mini-batch size of 30K. Even though the optimal number of warmup epochs appears to be linearly scaled with the mini-batch size and learning rate, it does not appear that the warmup epochs is scaled by k. The mini-batch size and learning rate from the second set of experiments are scaled by $k = 4$, yet the optimal number of warmup epochs increases by a factor of 2.

The number of iterations is $I = T/(s * k)$, where T is the total size of the training dataset and $(s * k)$ is the mini-batch size. Increasing mini-batch size reduces the number of iterations, which also means reducing the number weight updates. According to

Authors' addresses: Marie McCord, Middle Tennessee State University, USA, mam2hu@mtmail.mtsu.ed; Sajal Dash, NCCS, Oak Ridge National Laboratory, USA, dashs@ornl.gov; Feiyi Wang, NCCS, Oak Ridge National Laboratory, USA, ; Arjun Shankar, NCCS, Oak Ridge National Laboratory, USA, .

Manuscript submitted to ACM

linear rate scaling, a larger learning rate scaled by k that takes bigger steps is required to compensate for the fewer number of weight updates. It would be logical to assume that the incremental scale used for gradual warmup would also increase to take bigger steps. If that were the case, then a proportionally longer warmup period would be necessary to accommodate a larger desired learning rate, but the incremental scale would adjust as well to take larger steps. Then, the number of warmup epochs would be scaled by some amount less than k, instead of k itself. More experimentation is needed to determine the exact amount to scale the warmup epochs and the precise logic behind it.

To analyze linear rate scaling, we compared results from training with linear scaling to results from training without any scaling. We used a mini-batch size of 30,720 with 5 warmup epochs. Our results show that there is no significant difference between the experiment with linear scaling and the experiment without any scaling. Both achieve similar, subpar accuracies that are below the top-1 tier accuracy of 75.3%. This suggests that at large mini-batch sizes of 30k, linear rate scaling may have very little effect. A better scaling method for the learning rate could produce better results for mini-batch sizes this large.

Future work will continue to analyze techniques that successfully scale mini-batch size. Through understanding the range of parameters affected by each technique and the underlying training dynamics, we hope to create a generalizable strategy to increase mini-batch size while preserving accuracy.

ACM Reference Format:
Marie McCord, Sajal Dash, Feiyi Wang, and Arjun Shankar. 2021. Increasing Mini-Batch Size While Preserving Accuracy for Distributed Deep Learning. 1, 1 (October 2021), 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 ACKNOWLEDGEMENTS

This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

This work was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Science Undergraduate Laboratory Internship program.