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

Marie McCord, Middle Tennessee State University
mam2hu@mtmail.mtsu.edu

Sajal Dash, NCCS, Oak Ridge National Laboratory
dashs@ornl.gov

Feiyi Wang, NCCS, Oak Ridge National Laboratory

Arjun Shankar, NCCS, Oak Ridge National Laboratory

Introduction

Data-parallel distributed training among multiple workers reduces training time.

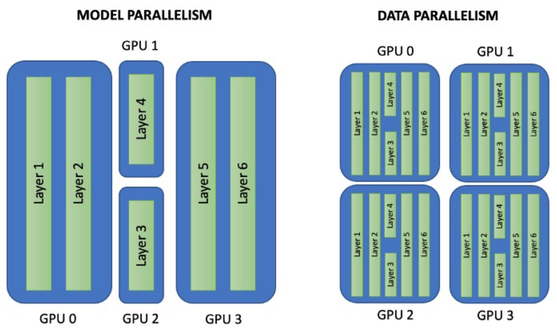


Figure 1: Distributed training techniques [1]

Large mini-batch sizes help maintain a balanced computation-to-communication ratio among workers. However, beyond a certain size limit, the benefits of large mini-batch sizes come at the cost of accuracy.

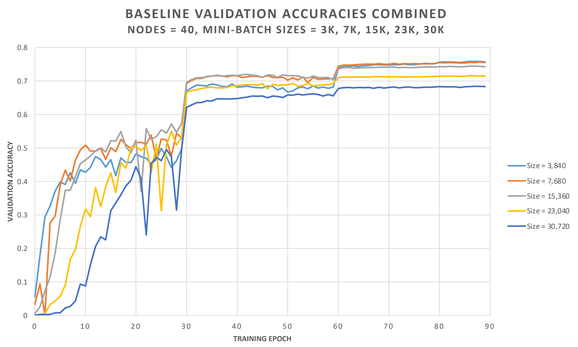


Figure 2: Accuracy decreases as mini-batch size increases.

Gradual warmup and *linear rate scaling* have been shown to scale mini-batch size to some extent. *Gradual warmup* slowly increases a very low learning rate to the desired rate during the first few epochs. *Linear rate scaling* linearly scales the learning rate with mini-batch size. This work analyzes both techniques.

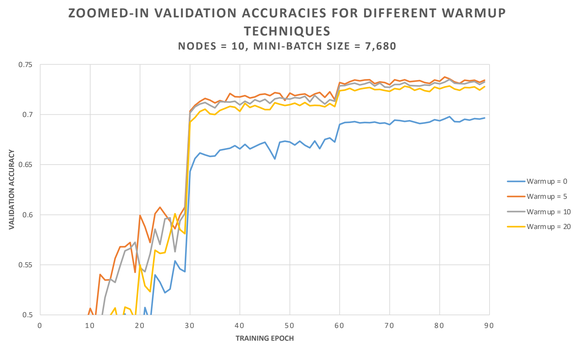
Methods

All experiments were run on Summit with the ResNet-50 model trained for 90 epochs on the ImageNet dataset.

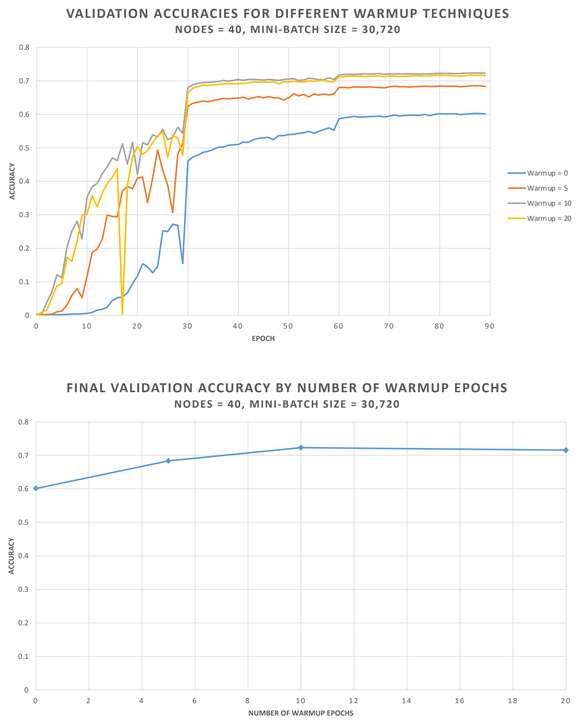
- Analyzed gradual warmup by training the model with 0, 5, 10, and 20 warmup epochs
- Analyzed linear rate scaling by comparing it to training without scaling of the base learning rate

Results

Gradual Warmup Technique:



Figures 3 & 4: No warmup epochs produces the worst accuracy rates. 5 warmup epochs is optimal for mini-batch size 7K. Accuracy decreases beyond 5 warmup epochs.



Figures 5 & 6: No warmup epochs produces the worst accuracy rates. 10 warmup epochs is optimal for mini-batch size 30K. Accuracy decreases beyond 10 warmup epochs.

Learning Rate Scaling:

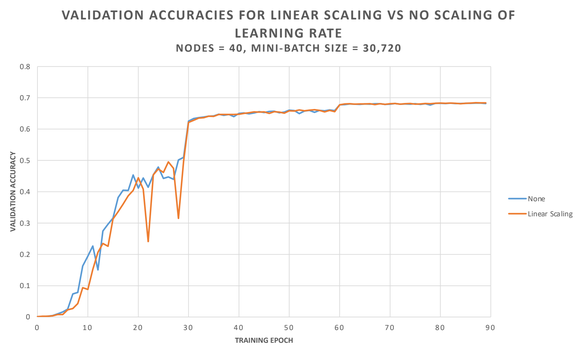


Figure 7: No significant difference between training with linear rate scaling and without scaling of base learning rate = 0.0125

Conclusions

- Gradual warmup is needed to stabilize early training phases, which are highly sensitive to learning rates after random weight initialization
- There is strong indication that the optimal number of warmup epochs is linearly scaled with mini-batch size
- This linear association is because larger mini-batch sizes reduce the number of training iterations, and fewer weight updates require a longer warmup period
- Gradual warmup beyond the optimal number of epochs reduces accuracy due to prolonging low learning rates that update weights too slowly
- There is no evidence linear rate scaling has a noticeable effect on mini-batch sizes as large as 30K

Scan QR code for paper →

References

[1] Torres.AI, J. (2020, November 23). [Model parallelism vs data parallelism]. Towards Data Science. https://towardsdatascience.com/scalable-deep-learning-on-parallel-and-distributed-infrastructures-e5fb4a956bef

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.

www.PosterPresentations.com