Literature

Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour

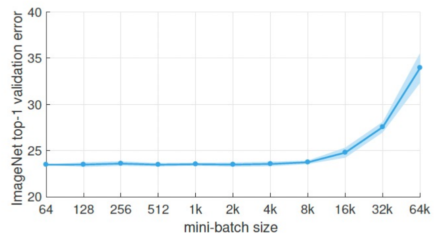
Priya Goyal Piotr Dollár Ross Girshick Pieter Noordhuis
Lukasz Wesolowski Aapo Kyrola Andrew Tulloch Yangqing Jia Kaimin

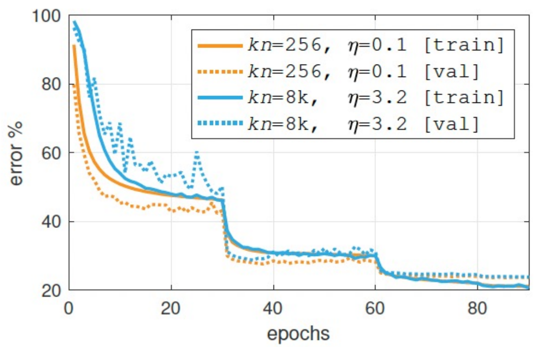
Facebook

.CV] 30 Apr 2018

Abstract

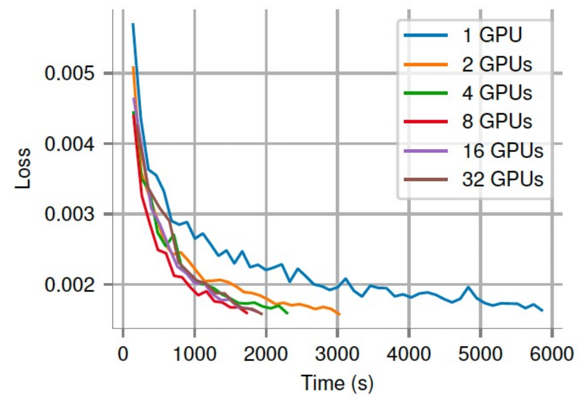
Deep learning thrives with large neural networks and large datasets. However, larger networks and larger datasets result in longer training times that impede research and development progress. Distributed synchronous SGD offers a potential solution to this problem by dividing SGD minibatches over a pool of parallel workers. Yet to make this scheme efficient, the per-worker workload must be large, which implies nontrivial growth in the SGD mini-batch size. In this paper we empirically show that on the

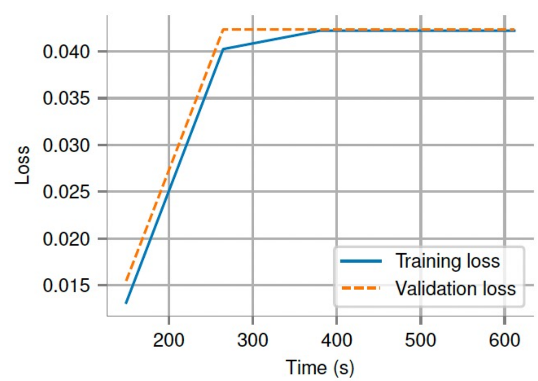




An argument in favor of strong scaling for deep neural networks with small datasets

Renato L. de F. Cunha, Eduardo R. Rodrigues, Matheus Palhares Viana, Dario Augusto Borges Oliveira
IBM Research





Weak scaling vs. strong scaling

Strong scaling

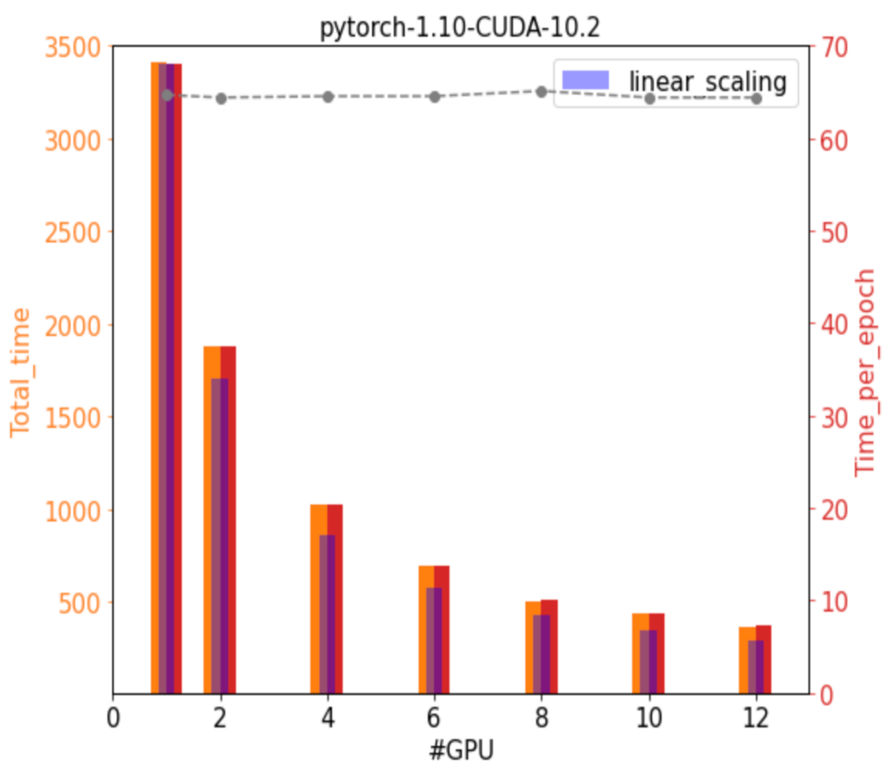
#GPUs	GPUs IDs	per-GPU batch size	effective batch size
1	0	1024	1024
2	0,1	512	1024
4	0,1,2,3	256	1024
8	0,1,2,3,4,5,6,7	128	1024

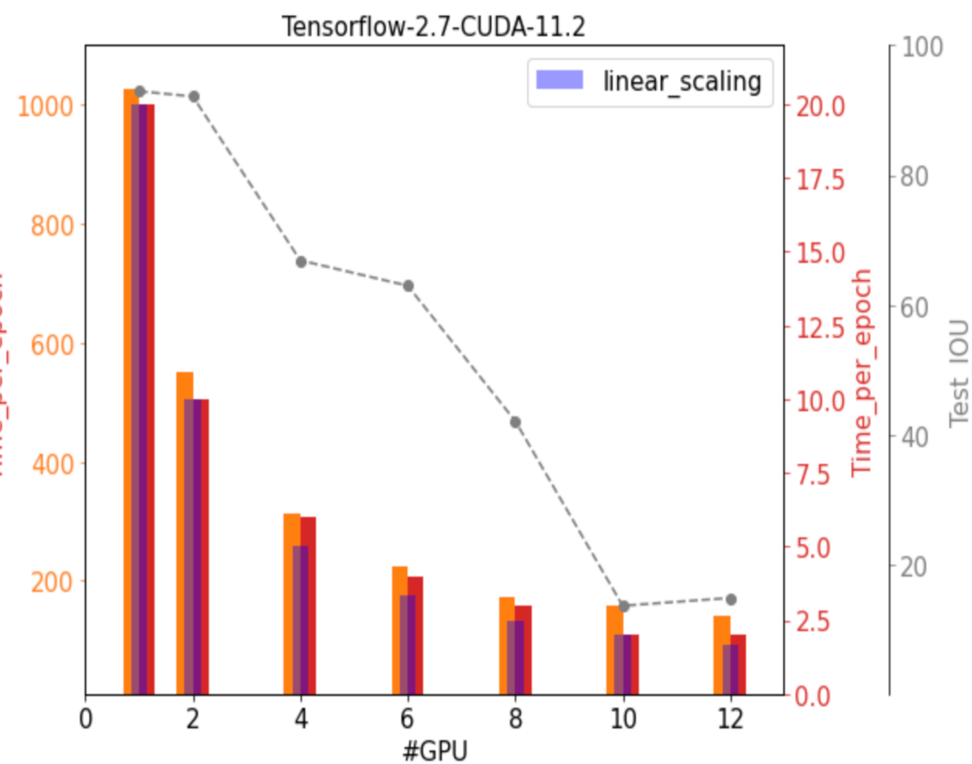
Weak scaling

#GPUs	GPUs IDs	per-GPU batch size	effective batch size
1	0	128	128
2	0,1	128	256
4	0,1,2,3	128	512
8	0,1,2,3,4,5,6,7	128	1024

Weak scaling

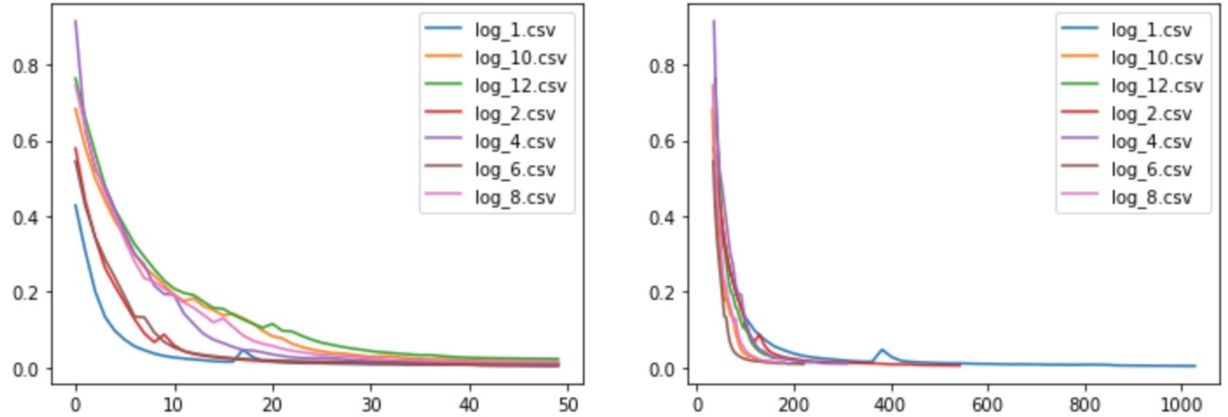
- Unet, data parallel, weak scaling: BS per process fixed:16, 12 cpus per job, 50 epochs
- Point of interest and emphasis: performance and/on small datasets



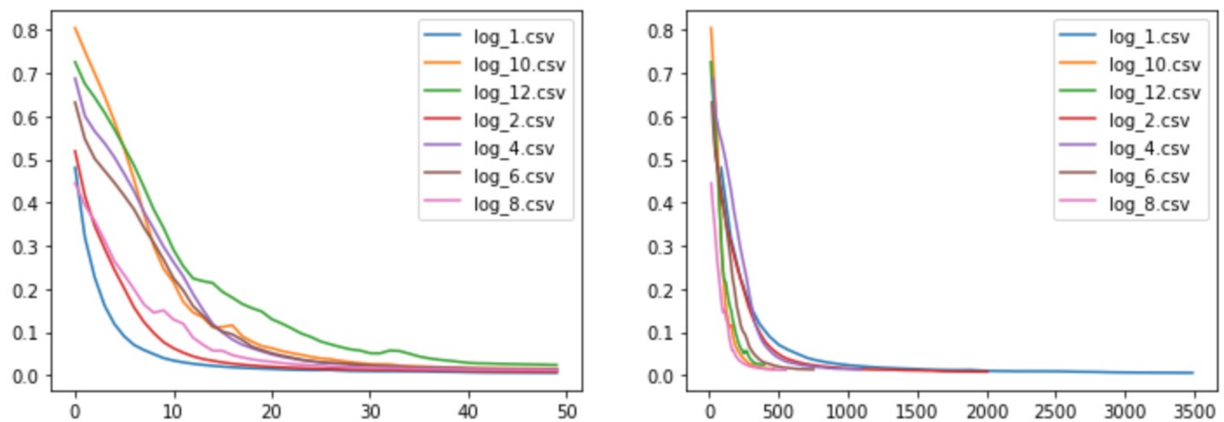


Train test loss: after linear scaling (+ warm up)

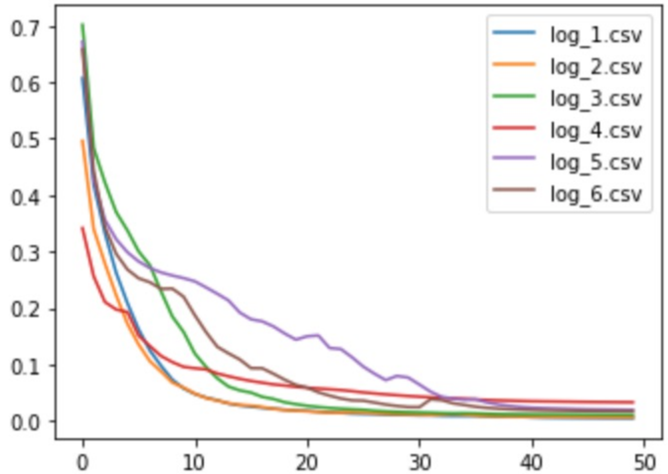
Tensorflow train loss: LS+ warm up for 10 epochs + scheduling

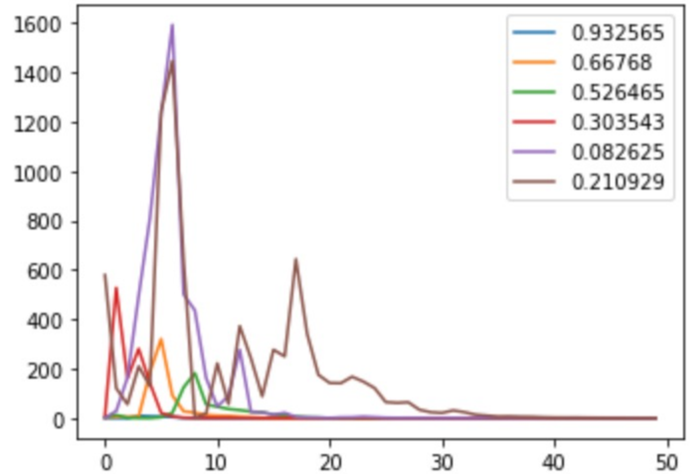


Pytorch train loss: LS+ warm up for 10 epochs + scheduling

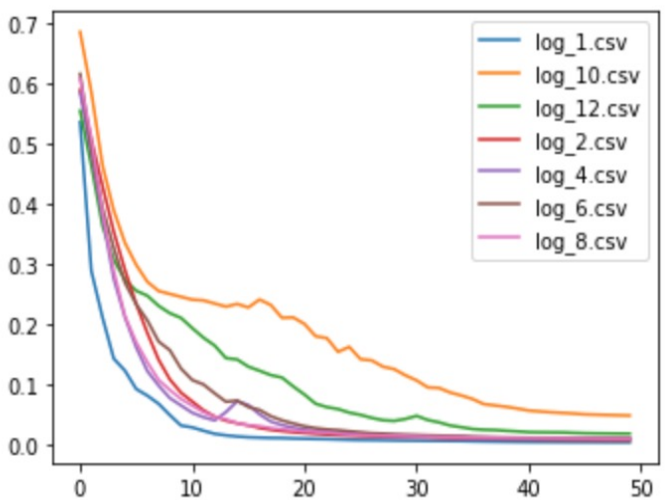


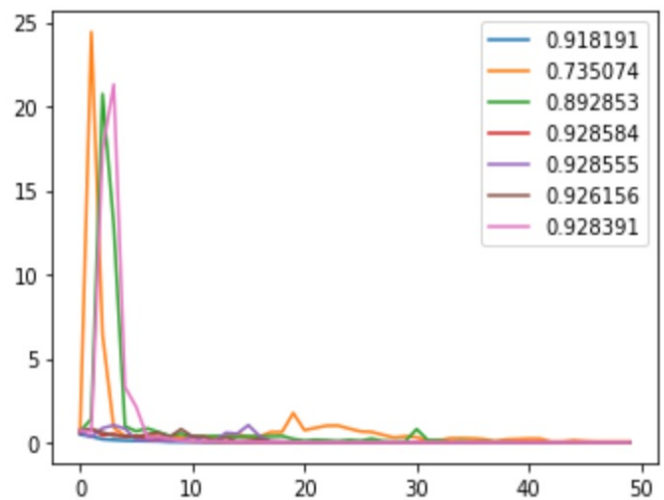
Tensorflow train and test loss: LS+ scheduling





Pytorch train and test loss: LS+ scheduling





Possible avenues:

1- A larger dataset for testing both strong and weak scaling effect on performance

2- Rule of thumb for when the performance falls

3- Horovod and a full comparative study