Fast and Accurate Deep Neural Network Training

Yang You
National University of Singapore

***Abstract*—In the last three years, supercomputers have become increasingly popular in leading AI companies. Amazon built a High Performance Computing (HPC) cloud. Google released its first 100-petaFlop supercomputer (TPU Pod). Facebook made a submission on the Top500 supercomputer list. Why do they like supercomputers? Because the computation of deep learning is very expensive. For example, even with 16 TPUs, BERT training takes more than 3 days. On the other hand, supercomputers can process 10^17 floating point operations per second. So why don't we just use supercomputers and finish the training of deep neural networks in a very short time? The reason is that deep learning does not have enough parallelism to make full use of thousands or even millions of processors in a typical modern supercomputer. There are two directions for parallelizing deep learning: model parallelism and data parallelism. Model parallelism is very limited. For data parallelism, current optimizers can not scale to thousands of processors because large-batch training is a sharp minimum problem. In this talk, I will introduce LARS (Layer-wise Adaptive Rate Scaling) and LAMB (Layer-wise Adaptive Moments for Batch training) optimizers, which can find more parallelism for deep learning. They can not only make deep learning systems scale well, but they can also help real-world applications to achieve higher accuracy.**

Since 2017, all the Imagenet training speed world records have been achieved using LARS. LARS was added to MLperf, which is the industry benchmark for fast deep learning. Google used LAMB to reduce BERT training time from 3 days to 76 minutes and achieve new state-of-the-art results on GLUE, RACE, and SQuAD benchmarks. The approaches introduced in this talk have been used by state-of-the-art distributed systems at Google, Intel, NVIDIA, Sony, Tencent, and so on.

BIOGRAPHY

Yang You is an assistant professor (tenure-track) at National University of Singapore. He received his PhD in Computer Science from UC Berkeley. His advisor is Prof. James Demmel, who was the former chair of the Computer Science Division and EECS Department. Yang You's research interests include Parallel/Distributed Algorithms, High Performance Computing, and Machine Learning. The focus of his current research is scaling up deep neural networks training on distributed systems or supercomputers. In 2017, his team broke the world record of ImageNet training speed, which was covered by the technology media like NSF, ScienceDaily, Science NewsLine, and i-programmer. In 2019, his team broke the world record of BERT training speed. The BERT training techniques have been used by many tech giants like Google, Microsoft, and NVIDIA. Yang You's LARS and LAMB optimizers are available in industry benchmark MLPerf. He is a winner of IPDPS 2015 Best Paper Award (0.8%), ICPP 2018 Best Paper Award (0.3%) and ACM/IEEE George Michael HPC Fellowship. Yang You is a Siebel Scholar and a winner of Lotfi A. Zadeh Prize. For more information, please check his lab's homepage at https://ai.comp.nus.edu.sg/