KTH Royal Institute of Technology
Stockholm

School of Electrical Engineering and
Computer Science

Scalable Machine Learning and Deep Learning - ID2223

Review Questions 5

Author
Emil Ståhl

Author
Erik Kongpachith

Author
Selemawit Fsha Nguse

December 11, 2021

1 Explain how does the data-parallelized learning work?

Data-parallellized learning is used in order to overcome the problem of training deep neural networks which is computational intensive and time consuming. Data-parallellized learning addresses the first issue with using massive amount of training dataset for training. This is done by replicating a whole model on multiple devices and using a different mini-batch for each. The gradient for each device is computed locally which is later aggregated to one gradient, this gradient is then used to update the weights of all devices.

2 Explain how does the model-parallelized learning work?

Model-parallellized learning is used in order to overcome the problem of training deep neural networks which is computational intensive and time consuming. Model-parallelized learning addresses the second issue with having a large number of parameters for a model. This is done by splitting a whole model across multiple devices. The splitting method used depends on the architecture of the neural network.

3 Explain different synchronization approaches in data-parallelized learning?

Synchronous

After each iteration, the workers synchronize their parameter updates. Every worker must wait for all workers to finish the transmission of all parameters in the current iteration, before the next training. Stragglers(slow workers) can influence the overall system throughput. A drawback with synchronous is high communication cost that limits the system scalability.

Stale-synchronous

Stale-synchronous addresses the straggler problem in synchronous. The faster workers do more updates than the slower workers to reduce the waiting time of the faster workers. Introduce a staleness bounded barrier to limit the iteration gap between the fastest worker and the slowest worker.

Asynchronous

There is no synchronization. Each worker transmits its gradients to the PS after it calculates the gradients. The PS updates the global model without waiting for the other workers.

Local SGD

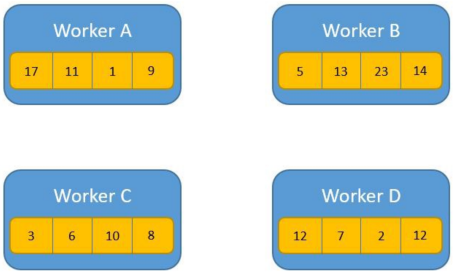
All workers run several iterations, and then averages all local models into the newest global model.

4 Briefly explain gradient quantization and gradient sparsification.

Gradient quantization and gradient sparsification is used for communication compression with the purpose of reducing the communication traffic in distributed deep learning while having little impact on the model convergence. The technique compresses the exchanged gradients or models before transmitting them across the network.

- Quantization
  - Uses lower bits to represent data
  - Gradients are of low precision
- Sparsification
  - Reducing the number of elements that are transmitted at each iteration.
  - Only significant gradients are required to update the model parameter to guarantee the convergence of the training.

5 Use the following picture and show step-by-step how the ring-allreduce works to compute the maximum of all elements?



Worker A	17	11	1	9
Worker B	5	13	23	14
Worker C	3	6	10	8
Worker D	12	7	2	12

Worker A		11	1	12
Worker B	17		23	14
Worker C	3	13		8
Worker D	12	7	10	

Worker A		11	10	
Worker B			23	14
Worker C	17			8
Worker D	12	13		

Worker A		13		
Worker B			23	
Worker C				14
Worker D	17			

Worker A	17	13		
Worker B		13	23	
Worker C			23	14
Worker D	17			14

Worker A	17	13		14
Worker B	17	13	23	
Worker C		13	23	14
Worker D	17		23	14

Worker A	17	13	23	14
Worker B	17	13	23	14
Worker C	17	13	23	14
Worker D	17	13	23	14