Review Questions 6

Group 1
Chuan Su
Diego Alonso Guillen Rosaperez

December 15, 2019

1. Both model parallelization and data parallelization are to training a single model across multiple devices. In model parallelization approach the model is split across multiple devices. Model parallelism depends on the architecture of the Neural Network. Data parallelization approach parallelizes the training of a neural network is to replicate it on each device, run a training step simultaneously on all replicas using a different mini-batch for each, and then aggregate the gradients to update the model parameters. .

2. **MultiWorkerMirroredStrategy**.implements synchronous distributed training across multiple workers, each with potentially multiple GPUs. Similar to MirroredStrategy, it creates copies of all variables in the model on each device across all workers.
ParameterServer supports parameter servers training on multiple machines. In this setup, some machines are designated as workers and some as parameter servers. Each variable of the model is placed on one parameter server. Computation is replicated across all GPUs of all the workers. For large models with millions of parameters, it is useful to shard these parameters across multiple parameter servers, to reduce the risk of saturating a single parameter servers network card.

3. Gradient Quantization and Gradient Sparsification are ways for compressing model updates, so as to reduce the communication overhead in data parallelization. More specifically Gradient Quantization is to reducing the number of bits per gradient whilst Gradient Sparsification is to communicating only important gradients that have a significant value.

4. Assuming sequence Worker A-¿Worker B-¿Worker D-¿Worker C-¿Worker A
Step 1: Share Reduce

   (a) For r0:
   A sends 17 to Worker B
   B applies reduce operator $17 + 5 = 22$ and sends result 22 to D
   D applies reduce operator $22 + 12 = 34$ and sends result 13 to C
   C applies reduce operator $34 + 3 = 37 = r0$

(b) For r1:
B sends 13 to Worker D
D applies reduce operator $13 + 7 = 20$ and sends result 20 to C
C applies reduce operator $20 + 6 = 26$ and sends result 26 to C
A .applies reduce operator $26 + 11 = 37 = r1$

(c) For r2:
D sends 2 to Worker C
C applies reduce operator $2 + 10 = 12$ and sends result 12 to A
A applies reduce operator $12 + 1 = 13$ and sends result 13 to B
B applies reduce operator $13 + 23 = 36 = r2$

(d) For r3:
C sends 8 to Worker A
A applies reduce operator $8 + 9 = 17$ and sends result 17 to B
B applies reduce operator $17 + 14 = 31$ and sends result 15 to D
D applies reduce operator $31 + 12 = 43 = r3$

Step 2: Share-only

(a) 1st iteration:
C sends $r0 = 37$ to A
A sends $r1 = 37$ to B
B sends $r2 = 36$ to D
D sends $r3 = 43$ to C

(b) 2nd iteration:
A sends $r0 = 37$ to B
B sends $r1 = 37$ to D
D sends $r2 = 36$ to C
C sends $r3 = 43$ to A

(c) 3nd iteration:
B sends $r0 = 37$ to D
D sends $r1 = 37$ to C
C sends $r2 = 36$ to A
A sends $r3 = 43$ to B

(d) At the end, each worker has the following elements: $37, 37, 36, 43$.