Scalable Machine Learning and Deep Learning - Review Questions 6

Deadline: December 15, 2019

1. What are model parallelism and data parallelism?

---

2. When training a model across multiple servers, what distribution strategies can you use? How do you choose which one to use?

---

3. Briefly explain gradient quantization and gradient sparsification.

---

4. Use the following picture and show step-by-step how the ring-allreduce works to compute the sum of all elements?