Topics in Machine Learning Systems

Author: Guanzhou (Jose) Hu 胡冠洲 @ UW-Madison CS839

Teacher: Prof. Shivaram

This note includes the list of paper we discussed this semester.

Compute

Background

- GPU Computing

  https://pages.cs.wisc.edu/~markhill/restricted/ieeemicro10_gpu.pdf
- cuDNN

  https://arxiv.org/pdf/1410.0759.pdf

Advances

- Autograd & JAX

  http://videolectures.net/deeplearning2017_johnson_automatic_differentiation/

  https://mlsys.org/Conferences/2019/doc/2018/146.pdf
- Rammer (rTasks)

  https://www.usenix.org/conference/osdi20/presentation/ma

Communication

Background

- Collective Communication

  https://www.cs.utexas.edu/~pingali/CSE392/2011sp/lectures/Conc_Comp.pdf

Advances

- BytePS

  https://www.usenix.org/conference/osdi20/presentation/jiang
- Horovod

  https://xiexbing.github.io/files/acodl_nsdi.pdf

- Megatron-LM

  https://arxiv.org/pdf/2104.04473.pdf

Profiling

Background

- Continuous Profiling

  https://dl.acm.org/doi/pdf/10.1145/268998.266637

- Magpie

  https://www.microsoft.com/en-us/research/wp-content/uploads/2003/05/magpiehotos03.pdf

Advances

- `nvprof`

  https://developer.nvidia.com/blog/cuda-pro-tip-nvprof-your-handy-universal-gpu-profiler/

- NCCL

  https://on-demand.gputechconf.com/gtc/2017/presentation/s7155-jeaugey-nccl.pdf

Serving

Background

- SEDA

  http://www.sosp.org/2001/papers/welsh.pdf

Advances

- Clockwork

  https://www.usenix.org/conference/osdi20/presentation/gujarati

- TVM & VTA

  https://arxiv.org/pdf/1807.04188.pdf

- HummingBird

  https://www.usenix.org/system/files/osdi20-nakandala.pdf

Scheduling

Background

- Omega

  https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/41684.pdf

Advances

- ASHA

  https://proceedings.mlsys.org/paper/2020/file/f4b9ec30ad9f68f89b29639786cb62ef-Supplemental.pdf

- Marius++

  https://arxiv.org/abs/2202.02365

- Gavel

  https://www.usenix.org/conference/osdi20/presentation/narayanan-deepak

Model Specific

- Deep Recommendation

  https://arxiv.org/pdf/2011.05497.pdf

- MHA & Transformers

https://proceedings.mlsys.org/paper/2021/hash/c9e1074f5b3f9fc8ea15d152add07294-Abstract.html

- Mixture-of-Experts (MoE)

  https://arxiv.org/abs/2201.05596

- Reinforcement (RL-Scope)

  https://proceedings.mlsys.org/paper/2021/file/d1fe173d08e959397adf34b1d77e88d7-Paper.pdf