

Reading List

- Sergeev, A. and Del Balso, M., 2018. Horovod: fast and easy distributed deep learning in TensorFlow. arXiv preprint arXiv:1802.05799.
- Shoeybi, M., Patwary, M., Puri, R., LeGresley, P., Casper, J. and Catanzaro, B., 2019. Megatron-Lm: Training multi-billion parameter language models using model parallelism. arXiv preprint arXiv:1909.08053.
- Jia, Z., Zaharia, M. and Aiken, A., 2019. Beyond Data and Model Parallelism for Deep Neural Networks. Proceedings of Machine Learning and Systems, 1, pp.1-13.
- Narayanan, D., Harlap, A., Phanishayee, A., Seshadri, V., Devanur, N.R., Ganger, G.R., Gibbons, P.B. and Zaharia, M., 2019, October. PipeDream: generalized pipeline parallelism for DNN training. In Proceedings of the 27th ACM Symposium on Operating Systems Principles (pp. 1-15).
- Gholami, A., Azad, A., Jin, P., Keutzer, K. and Buluc, A., 2018, July. Integrated model, batch, and domain parallelism in training neural networks. In Proceedings of the 30th on Symposium on Parallelism in Algorithms and Architectures (pp. 77-86).
- Ben-Nun, T. and Hoefler, T., 2019. Demystifying parallel and distributed deep learning: An in-depth concurrency analysis. ACM Computing Surveys (CSUR), 52(4), pp.1-43.