REFERENCES

[1] A. F. Agarap, *Training Deep Neural Networks for Image Classification in a Homogenous Distributed System*. 2019. doi: 10.13140/RG.2.2.21940.40321.

[2] H. Xu *et al.*, "Compressed Communication for Distributed Deep Learning: Survey and Quantitative Evaluation," 2020, Accessed: Aug. 25, 2023. [Online]. Available: https://repository.kaust.edu.sa/handle/10754/662495

[3] J. Guo *et al.*, *Accelerating Distributed Deep Learning By Adaptive Gradient Quantization*. 2020, p. 1607. doi: 10.1109/ICASSP40776.2020.9054164.

[4] R. Zhang *et al.*, "LLaMA-Adapter: Efficient Fine-tuning of Language Models with Zero-init Attention." arXiv, Jun. 14, 2023. doi: 10.48550/arXiv.2303.16199.

[5] J. Dean *et al.*, "Large Scale Distributed Deep Networks," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2012. Accessed: Aug. 28, 2023. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2012/hash/6aca97005c68f1206823815f66102863-Abstract.html

[6] H. Mahmud, P. Kang, K. Desai, P. Lama, and S. Prasad, "A Converting Autoencoder Toward Low-latency and Energy-efficient DNN Inference at the Edge." arXiv, Mar. 11, 2024. doi: 10.48550/arXiv.2403.07036.

[7] B. Hasheminezhad, S. Shirzad, N. Wu, P. Diehl, H. Schulz, and H. Kaiser, *Towards a Scalable and Distributed Infrastructure for Deep Learning Applications*. 2020.

[8] R. Mayer and H.-A. Jacobsen, "Scalable Deep Learning on Distributed Infrastructures: Challenges, Techniques and Tools," no. arXiv:1903.11314. arXiv, Sep. 25, 2019. doi: 10.48550/arXiv.1903.11314.

[9] N. Mungoli, "Scalable, Distributed AI Frameworks: Leveraging Cloud Computing for Enhanced Deep Learning Performance and Efficiency," no. arXiv:2304.13738. arXiv, Apr. 26, 2023. Accessed: Aug. 22, 2023. [Online]. Available: http://arxiv.org/abs/2304.13738

[10] S. Shi, Q. Wang, and X. Chu, "Performance Modeling and Evaluation of Distributed Deep Learning Frameworks on GPUs," no. arXiv:1711.05979. arXiv, Aug. 20, 2018. Accessed: Aug. 28, 2023. [Online]. Available: http://arxiv.org/abs/1711.05979

[11] "(PDF) Performance Analysis and Comparison of Distributed Training Strategies for Deep Learning." Accessed: Aug. 26, 2023. [Online]. Available: https://www.researchgate.net/publication/358552709_Performance_Analysis_and_Comparison_of_Distributed_Training_Strategies_for_Deep_Learning

[12] "Sensors | Free Full-Text | Analysis of the Application Efficiency of TensorFlow and PyTorch in Convolutional Neural Network." Accessed: Feb. 28, 2024. [Online]. Available: https://www.mdpi.com/1424-8220/22/22/8872

[13] "(PDF) Performance Analysis of Data Parallelism Technique in Machine Learning for Human Activity Recognition Using LSTM." Accessed: Aug. 26, 2023. [Online]. Available: https://www.researchgate.net/publication/338946032_Performance_Analysis_of

_Data_Parallelism_Technique_in_Machine_Learning_for_Human_Activity_Recognition_Using_LSTM

[14] J.-S. Lerat, S. A. Mahmoudi, and S. Mahmoudi, “Distributed Deep Learning: From Single-Node to Multi-Node Architecture,” *Electronics*, vol. 11, no. 10, Art. no. 10, Jan. 2022, doi: 10.3390/electronics11101525.

[15] M. Migliorini, R. Castellotti, L. Canali, and M. Zanetti, “Machine Learning Pipelines with Modern Big Data Tools for High Energy Physics,” *Comput. Softw. Big Sci.*, vol. 4, no. 1, Art. no. 1, Dec. 2020, doi: 10.1007/s41781-020-00040-0.

[16] M. Bień, *Big Data Analysis and Machine Learning at Scale with Oracle Cloud Infrastructure*. 2019. doi: 10.5281/zenodo.3550777.

[17] M. H. Anwar, S. Ghafouri, S. S. Gill, and J. Doyle, “Recommender System for Optimal Distributed Deep Learning in Cloud Datacenters,” *Wirel. Pers. Commun.*, vol. 127, no. 2, Art. no. 2, Nov. 2022, doi: 10.1007/s11277-021-08699-3.

[18] Z. Tang, S. Shi, X. Chu, W. Wang, and B. Li, “Communication-Efficient Distributed Deep Learning: A Comprehensive Survey.” arXiv, Mar. 10, 2020. doi: 10.48550/arXiv.2003.06307.

[19] A. Zam, “Evaluating Distributed Machine Learning using IoT Devices”.

[20] G. Xiao, J. Lin, M. Seznec, H. Wu, J. Demouth, and S. Han, “SmoothQuant: Accurate and Efficient Post-Training Quantization for Large Language Models,” in *Proceedings of the 40th International Conference on Machine Learning*, PMLR, Jul. 2023, pp. 38087–38099. Accessed: Aug. 28, 2023. [Online]. Available: https://proceedings.mlr.press/v202/xiao23c.html

[21] A. Borzunov *et al.*, “Distributed Inference and Fine-tuning of Large Language Models Over The Internet,” *Adv. Neural Inf. Process. Syst.*, vol. 36, pp. 12312–12331, Dec. 2023.

[22] “tf.distribute.MultiWorkerMirroredStrategy | TensorFlow v2.16.1.” Accessed: Apr. 27, 2024. [Online]. Available: https://www.tensorflow.org/api_docs/python/tf/distribute/MultiWorkerMirroredStrategy