





















## REFERENCES

- [1] Mahmoud Assran, Nicolas Loizou, Nicolas Ballas, and Mike Rabbat. 2019. Stochastic gradient push for distributed deep learning. In *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, 344–353.
- [2] Michael Blot, David Picard, and Matthieu Cord. 2018. GoSGD: Distributed Optimization for Deep Learning with Gossip Exchange. *arXiv preprint arXiv:1804.01852* (2018).
- [3] Michael Blot, David Picard, Matthieu Cord, and Nicolas Thome. 2016. Gossip training for deep learning. *Proceedings of the Advances in Neural Information Processing Systems Workshop on Optimization for Machine Learning* (2016).
- [4] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020).
- [5] Trishul Chilimbi, Yutaka Suzue, Johnson Apacible, and Karthik Kalyanaraman. 2014. Project adam: Building an efficient and scalable deep learning training system. In *Proceedings of the Symposium on Operating Systems Design and Implementation (OSDI)*, 571–582.
- [6] Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Mao, Andrew Senior, Paul Tucker, Ke Yang, Quoc V Le, et al. 2012. Large scale distributed deep networks. In *Proceedings of the Advances in Neural Information Processing Systems*. 1223–1231.
- [7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
- [8] Nima Eshraghi and Ben Liang. 2020. Distributed Online Optimization over a Heterogeneous Network with Any-Batch Mirror Descent. In *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, 2933–2942.
- [9] Alexandros V Gerbessiotis and Leslie G Valiant. 1994. Direct bulk-synchronous parallel algorithms. *J. Parallel and Distrib. Comput.* 22, 2 (1994), 251–267.
- [10] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. 2017. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677* (2017).
- [11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778.
- [12] Qirong Ho, James Cipar, Henggang Cui, Seunghak Lee, Jin Kyu Kim, Phillip B Gibbons, Garth A Gibson, Greg Ganger, and Eric P Xing. 2013. More effective distributed ml via a stale synchronous parallel parameter server. In *Proceedings of the Advances in Neural Information Processing Systems*. 1223–1231.
- [13] Shlomo Hoory, Nathan Linial, and Avi Wigderson. 2006. Expander graphs and their applications. *Bull. Amer. Math. Soc.* 43, 4 (2006), 439–561.
- [14] Yanping Huang, Youlong Cheng, Ankur Bapna, Orhan Firat, Dehao Chen, Mia Chen, HyoukJoong Lee, Jiquan Ngiam, Quoc V Le, Yonghui Wu, et al. 2019. Gpipe: Efficient training of giant neural networks using pipeline parallelism. In *Proceedings of the Advances in Neural Information Processing Systems*. 103–112.
- [15] Jiawei Jiang, Bin Cui, Ce Zhang, and Lele Yu. 2017. Heterogeneity-aware distributed parameter servers. In *Proceedings of the ACM International Conference on Management of Data, (SIGMOD)*. 463–478.
- [16] Tyler Johnson, Pulkit Agrawal, Haijie Gu, and Carlos Guestrin. 2020. AdaScale SGD: A User-Friendly Algorithm for Distributed Training. In *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, 4911–4920.
- [17] David Kempe, Alin Dobra, and Johannes Gehrke. 2003. Gossip-based computation of aggregate information. In *Proceedings of the IEEE Symposium on Foundations of Computer Science (FOCS)*. 482–491.
- [18] Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images. (2009).
- [19] Mu Li, David G Andersen, Jun Woo Park, Alexander J Smola, Amr Ahmed, Vanja Josifovski, James Long, Eugene J Shekita, and Bor-Yiing Su. 2014. Scaling distributed machine learning with the parameter server. In *Proceedings of the Symposium on Operating Systems Design and Implementation (OSDI)*. 583–598.
- [20] Youjie Li, Mingchao Yu, Songze Li, Salman Avestimehr, Nam Sung Kim, and Alexander Schwing. 2018. Pipe-SGD: a decentralized pipelined SGD framework for distributed deep net training. In *Proceedings of the International Conference on Neural Information Processing Systems (NIPS)*. 8056–8067.
- [21] Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. 2017. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. In *Proceedings of the Advances in Neural Information Processing Systems*. 5330–5340.
- [22] Xiangru Lian, Wei Zhang, Ce Zhang, and Ji Liu. 2018. Asynchronous decentralized parallel stochastic gradient descent. In *Proceedings of the international Conference on Machine Learning (ICML)*. 3049–3058.
- [23] Giorgi Nadiradze, Iliia Markov, Bapi Chatterjee, Vyacheslav Kungurtsev, and Dan Alistarh. 2021. Elastic Consistency: A Practical Consistency Model for Distributed Stochastic Gradient Descent. (2021).
- [24] Deepak Narayanan, Aaron Harlap, Amar Phanishayee, Vivek Seshadri, Nikhil R Devanur, Gregory R Ganger, Phillip B Gibbons, and Matei Zaharia. 2019. PipeDream: generalized pipeline parallelism for DNN training. In *Proceedings of the ACM Symposium on Operating Systems Principles (SOSP)*. 1–15.
- [25] Benjamin Recht, Christopher Re, Stephen Wright, and Feng Niu. 2011. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*. 693–701.
- [26] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115, 3 (2015), 211–252. <https://doi.org/10.1007/s11263-015-0816-y>
- [27] Alexander Sergeev and Mike Del Balso. 2018. Horovod: fast and easy distributed deep learning in TensorFlow. *arXiv preprint arXiv:1802.05799* (2018).
- [28] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- [29] Jianyu Wang and Gauri Joshi. 2018. Cooperative SGD: A unified framework for the design and analysis of communication-efficient SGD algorithms. *arXiv preprint arXiv:1808.07576* (2018).
- [30] Shaoqi Wang, Aidi Pi, and Xiaobo Zhou. 2019. Scalable distributed dl training: Batching communication and computation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, Vol. 33. 5289–5296.
- [31] Ran Xin, Soumya Kar, and Usman A Khan. 2020. Decentralized stochastic optimization and machine learning: A unified variance-reduction framework for robust performance and fast convergence. *IEEE Signal Processing Magazine* 37, 3 (2020), 102–113.
- [32] Yang You, Igor Gitman, and Boris Ginsburg. 2017. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888* (2017).
- [33] Yang You, Jonathan Hseu, Chris Ying, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. 2019. Large-batch training for LSTM and beyond. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. 1–16.
- [34] Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. 2019. Large batch optimization for deep learning: Training bert in 76 minutes. *arXiv preprint arXiv:1904.00962* (2019).
- [35] Chen Yu, Hanlin Tang, Cedric Renggli, Simon Kassing, Ankit Singla, Dan Alistarh, Ce Zhang, and Ji Liu. 2019. Distributed learning over unreliable networks. In *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, 7202–7212.
- [36] Hao Zhang, Zeyu Zheng, Shizhen Xu, Wei Dai, Qirong Ho, Xiaodan Liang, Zhiting Hu, Jinliang Wei, Pengtao Xie, and Eric P Xing. 2017. Poseidon: An Efficient Communication Architecture for Distributed Deep Learning on GPU Clusters. In *Proceedings of the USENIX Annual Technical Conference (ATC)*. 181–193.
- [37] Sixin Zhang, Anna E Choromanska, and Yann LeCun. 2015. Deep learning with elastic averaging SGD. In *Proceedings of the Advances in Neural Information Processing Systems*. 685–693.
- [38] Xing Zhao, Aijun An, Junfeng Liu, and Bao Xin Chen. 2019. Dynamic stale synchronous parallel distributed training for deep learning. In *Proceedings of the IEEE International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 1507–1517.
- [39] Xing Zhao, Manos Papagelis, Aijun An, Bao Xin Chen, Junfeng Liu, and Yonggang Hu. 2019. Elastic Bulk Synchronous Parallel Model for Distributed Deep Learning. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*. IEEE, 1504–1509.
- [40] Zhengyuan Zhou, Panayotis Mertikopoulos, Nicholas Bambos, Peter Glynn, Yinyu Ye, Li-Jia Li, and Li Fei-Fei. 2018. Distributed Asynchronous Optimization with Unbounded Delays: How Slow Can You Go?. In *Proceedings of the International Conference on Machine Learning (ICML)*. 5970–5979.