CS6285: Topics in Computer Science (Bridging System and Deep Learning)

1. Module Description

Co-design of system and machine learning algorithms has led to faster and more scalable machine learning systems. The module aims to expose students to recent state-of-the-art co-design techniques to make deep learning run faster, touching on both system research and AI research. The specific topics include distributed deep learning, large-batch training, second-order optimization, asynchronous algorithms, neural network compression, federated machine learning, neural networks pipeline processing, memory-efficient optimization, model parallelism, and efficient communication methods.

2. Pre-requisite(s)

CS5242 Neural Networks and Deep Learning (not required for graduate students)
CS3210 Parallel Computing (not required for graduate students)

3. Tentative Schedule

Week 01 Contents (by Yang You)
Introduction, high-performance system, parallel & distributed deep learning
Week 01 Readings

- [required] Introduction to Parallel Computing, Blaise Barney, Lawrence Livermore National Laboratory https://computing.llnl.gov/tutorials/parallel_comp/
- [optional] Introduction to Message Passing Interface (MPI), Blaise Barney, Lawrence Livermore National Laboratory https://computing.llnl.gov/tutorials/mpi/
- [required] Dean J, Corrado G, Monga R, Chen K, Devin M, Mao M, Ranzato MA, Senior A, Tucker P, Yang K, Le QV. Large scale distributed deep networks. In Advances in neural information processing systems 2012 (pp. 1223-1231).

Week 01 Homework [optional]

- Write a simple C++ neural network training by backpropagation.
- Do model-parallelism and data parallelism by MPI.
- https://github.com/huangzehao/SimpleNeuralNetwork

Week 02 Contents (by Yang You)
Deep learning optimizers and convergence

Week 02 Readings

- [optional] Ruder, Sebastian. "An overview of gradient descent optimization algorithms." arXiv preprint arXiv:1609.04747 (2016).
- [required] Chapter 8 of Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. Deep learning. MIT press, 2016.
    - https://www.deeplearningbook.org/contents/optimization.html
- [optional] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).
- [optional] Reddi, Sashank J., Satyen Kale, and Sanjiv Kumar. "On the convergence of adam and beyond." arXiv preprint arXiv:1904.09237 (2019).

Week 02 Homework [optional]

- Implement Momentum SGD, AdaGrad, Adam from scratch.
- Use your optimizers in your C++ backpropagation.

Week 03 Contents (by Yang You)

Large-Batch Optimization

Week 03 Readings

- [required] Keskar NS, Mudigere D, Nocedal J, Smelyanskiy M, Tang PT. On large-batch training for deep learning: Generalization gap and sharp minima. arXiv preprint arXiv:1609.04836. 2016 Sep 15.
- [required] You Y, Li J, Reddi S, Hseu J, Kumar S, Bhojanapalli S, Song X, Demmel J, Keutzer K, Hsieh CJ. Large batch optimization for deep learning: Training bert in 76 minutes. InInternational Conference on Learning Representations 2019 Sep 25.
- [optional] Krizhevsky A. One weird trick for parallelizing convolutional neural networks. arXiv preprint arXiv:1404.5997. 2014 Apr 23.
- [optional] Goyal P, Dollár P, Girshick R, Noordhuis P, Wesolowski L, Kyrola A, Tulloch A, Jia Y, He K. Accurate, large minibatch sgd: Training imagenet in 1 hour. arXiv preprint arXiv:1706.02677. 2017 Jun 8.
- [optional] You Y, Gitman I, Ginsburg B. Large batch training of convolutional networks. arXiv preprint arXiv:1708.03888. 2017 Aug 13.
- [optional] Jia X, Song S, He W, Wang Y, Rong H, Zhou F, Xie L, Guo Z, Yang Y, Yu L, Chen T. Highly scalable deep learning training system with mixed-precision: Training imagenet in four minutes. arXiv preprint arXiv:1807.11205. 2018 Jul 30.
- [optional] Ying C, Kumar S, Chen D, Wang T, Cheng Y. Image classification at supercomputer scale. arXiv preprint arXiv:1811.06992. 2018 Nov 16.
- [optional] Kumar S, Bitorff V, Chen D, Chou C, Hechtman B, Lee H, Kumar N, Mattson P, Wang S, Wang T, Xu Y. Scale MLPerf-0.6 models on Google TPU-v3 Pods. arXiv preprint arXiv:1909.09756. 2019 Sep 21.

Week 04 Contents (by students or Yang You)
Recent Optimization Techniques
Week 04 Readings

- [required] Loshchilov I, Hutter F. Decoupled Weight Decay Regularization. InInternational Conference on Learning Representations 2018 Sep 27.
- [required] Santurkar, S., Tsipras, D., Ilyas, A. and Madry, A., 2018. How does batch normalization help optimization?. In Advances in Neural Information Processing Systems (pp. 2483-2493).
- [optional] Miyato, T., Kataoka, T., Koyama, M. and Yoshida, Y., 2018, February. Spectral Normalization for Generative Adversarial Networks. In International Conference on Learning Representations.
- [optional] Liu L, Jiang H, He P, Chen W, Liu X, Gao J, Han J. On the variance of the adaptive learning rate and beyond. arXiv preprint arXiv:1908.03265. 2019 Aug 8.
- [optional] Luo L, Xiong Y, Liu Y, Sun X. Adaptive gradient methods with dynamic bound of learning rate. arXiv preprint arXiv:1902.09843. 2019 Feb 26.
- [optional] Zhang M, Lucas J, Ba J, Hinton GE. Lookahead Optimizer: k steps forward, 1 step back. InAdvances in Neural Information Processing Systems 2019 (pp. 9593-9604).

Week 05 Contents (guest lecture by Gregory Pauloski or students)
Second-order optimization and its distributed implementation
Week 05 Readings

- [required] Martens, James, and Roger Grosse. "Optimizing neural networks with kronecker-factored approximate curvature." In International conference on machine learning, pp. 2408-2417. 2015.
- [required] Pauloski, J. Gregory, Zhao Zhang, Lei Huang, Weijia Xu, and Ian T. Foster. "Convolutional Neural Network Training with Distributed K-FAC." arXiv preprint arXiv:2007.00784 (2020).
- [optional] Osawa K, Tsuji Y, Ueno Y, Naruse A, Yokota R, Matsuoka S. Large-scale distributed second-order optimization using kronecker-factored approximate curvature for deep convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2019 (pp. 12359-12367).

Week 06 Contents (by students or Yang You)
Asynchronous algorithms: parameter server, Hogwild, EA-SGD, federated learning
Week 06 Readings

- [optional] Recht B, Re C, Wright S, Niu F. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. InAdvances in neural information processing systems 2011 (pp. 693-701).
- [required] Li, Mu, David G. Andersen, Jun Woo Park, Alexander J. Smola, Amr Ahmed, Vanja Josifovski, James Long, Eugene J. Shekita, and Bor-Yiing Su. "Scaling distributed machine learning with the parameter server." In 11th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 14), pp. 583-598. 2014.
- [optional] Zhang S, Choromanska AE, LeCun Y. Deep learning with elastic averaging SGD. InAdvances in neural information processing systems 2015 (pp. 685-693).
- [required] Yang Q, Liu Y, Chen T, Tong Y. Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology (TIST). 2019 Jan 28;10(2):1-9.

Week 07 Contents (by students or Yang You)
Neural network compression: deep compression, deep gradient compression, deep leakage from gradients
Week 07 Readings

- [required] Han S, Mao H, Dally WJ. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. arXiv preprint arXiv:1510.00149. 2015 Oct 1.
- [required] Lin Y, Han S, Mao H, Wang Y, Dally WJ. Deep gradient compression: Reducing the communication bandwidth for distributed training. arXiv preprint arXiv:1712.01887. 2017 Dec 5.
- [optional] Wen, W., Xu, C., Yan, F., Wu, C., Wang, Y., Chen, Y. and Li, H., 2017. Terngrad: Ternary gradients to reduce communication in distributed deep learning. In Advances in neural information processing systems (pp. 1509-1519).
- [optional] Alistarh, D., Grubic, D., Li, J., Tomioka, R. and Vojnovic, M., 2017. QSGD: Communication-efficient SGD via gradient quantization and encoding. In Advances in Neural Information Processing Systems (pp. 1709-1720).
- [optional] Zhu L, Liu Z, Han S. Deep leakage from gradients. In Advances in Neural Information Processing Systems 2019 (pp. 14747-14756).

Week 08 Contents (by students or Yang You)
Memory-Efficient Optimization
Week 08 Readings

- [required] Shazeer, Noam, and Mitchell Stern. "Adafactor: Adaptive learning rates with sublinear memory cost." arXiv preprint arXiv:1804.04235 (2018).

- [required] Ginsburg B, Castonguay P, Hrinchuk O, Kuchaiev O, Lavrukhin V, Leary R, Li J, Nguyen H, Zhang Y, Cohen JM. Stochastic gradient methods with layer-wise adaptive moments for training of deep networks. arXiv preprint arXiv:1905.11286. 2019 May 27.
- [optional] Anil, Rohan, Vineet Gupta, Tomer Koren, and Yoram Singer. "Memory Efficient Adaptive Optimization." In Advances in Neural Information Processing Systems, pp. 9749-9758. 2019.
- [optional] Gupta, Vineet, Tomer Koren, and Yoram Singer. "Shampoo: Preconditioned stochastic tensor optimization." arXiv preprint arXiv:1802.09568 (2018).

Week 09 Contents (by students or Yang You)
Deep Neural Networks Pipeline Processing
Week 09 Readings

- [required] Huang Y, Cheng Y, Bapna A, Firat O, Chen D, Chen M, Lee H, Ngiam J, Le QV, Wu Y. Gpipe: Efficient training of giant neural networks using pipeline parallelism. InAdvances in Neural Information Processing Systems 2019 (pp. 103-112).
- [required] Chen, Tianqi, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. "Training deep nets with sublinear memory cost." arXiv preprint arXiv:1604.06174 (2016).
- [optional] Narayanan, Deepak, Aaron Harlap, Amar Phanishayee, Vivek Seshadri, Nikhil R. Devanur, Gregory R. Ganger, Phillip B. Gibbons, and Matei Zaharia. "PipeDream: generalized pipeline parallelism for DNN training." In Proceedings of the 27th ACM Symposium on Operating Systems Principles, pp. 1-15. 2019.

Week 10 Contents (by students or Yang You)
Efficient communication algorithm and system
Week 10 Readings

- [required] Patarasuk P, Yuan X. Bandwidth optimal all-reduce algorithms for clusters of workstations. Journal of Parallel and Distributed Computing. 2009 Feb 1;69(2):117-24.
- [required] Sergeev, Alexander, and Mike Del Balso. "Horovod: fast and easy distributed deep learning in TensorFlow." arXiv preprint arXiv:1802.05799 (2018).
- [optional] Bringing HPC Techniques to Deep Learning, Andrew Gibiansky, https://andrew.gibiansky.com/blog/machine-learning/baidu-allreduce/
- [optional] Jia X, Song S, He W, Wang Y, Rong H, Zhou F, Xie L, Guo Z, Yang Y, Yu L, Chen T. Highly scalable deep learning training system with mixed-precision: Training imagenet in four minutes. arXiv preprint arXiv:1807.11205. 2018 Jul 30.

Week 11 Contents (by students or Yang You)
Model Parallelism and Memory-Efficient System
Week 11 Readings

- [required] Shoeybi M, Patwary M, Puri R, LeGresley P, Casper J, Catanzaro B. Megatron-lm: Training multi-billion parameter language models using model parallelism. arXiv preprint arXiv:1909.08053. 2019 Sep.
- [required] Rajbhandari S, Rasley J, Ruwase O, He Y. ZeRO: Memory Optimization Towards Training A Trillion Parameter Models. arXiv preprint arXiv:1910.02054. 2019 Oct 4.
- [optional] Shazeer, Noam, Youlong Cheng, Niki Parmar, Dustin Tran, Ashish Vaswani, Penporn Koanantakool, Peter Hawkins et al. "Mesh-tensorflow: Deep learning for supercomputers." In Advances in Neural Information Processing Systems, pp. 10414-10423. 2018.

Week 12 Contents (by students or Yang You)
Real deep learning systems
Week 12 Readings

- [required] Lepikhin, Dmitry, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. "Gshard: Scaling giant models with conditional computation and automatic sharding." arXiv preprint arXiv:2006.16668 (2020).
- [required] Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan et al. "Language models are few-shot learners." arXiv preprint arXiv:2005.14165 (2020).
- [optional] Chen, Tianqi, Thierry Moreau, Ziheng Jiang, Lianmin Zheng, Eddie Yan, Haichen Shen, Meghan Cowan et al. "{TVM}: An automated end-to-end optimizing compiler for deep learning." In 13th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 18), pp. 578-594. 2018.

Week 13 Contents (by students)
Presentation for the final project
Week 13 Readings

- [optional] Patterson, David A. "How to give a bad talk."
    - https://drive.google.com/file/d/0Bzis5MXW83vCdUdXYnFIVDVOSkE/view
    - https://www.youtube.com/watch?v=Pbdo-ozuOug&feature=youtu.be

4. Evaluation and Grading

Each student writes a paper review per week (30%)
Each team finishes a paper presentation (20%)
Final project presentation (10%)
Final Project (40%)

The workload can be reduced (depending on the feedback from the students)

- Each student picks a paper to write a weekly paper review
    - Summary of the paper (a minimum of 250 words)
    - Strength and Weakness of the paper (a minimum of 250 words)
    - Future work based on this paper (a minimum of 100 words)
    - The deadline is 9:59pm, Sunday, each week (Singapore time)
        - Not required for week 01, week 02, and week 13
- Each team should finish an one-page proposal before the end of the 8th week
- The paper presentation should be around 90 minutes (1 or 2 papers)
- The final presentation can be 5 ~ 20 minutes (depending on the class size)

5. Module Information

- Time: Tuesday 3-5pm
- Location: zoom (due to COVID-19)

6. Instructor

Yang You
https://www.cs.berkeley.edu/~youyang/