Research Statement

Sunwoo Lee
(sunwoolee1.2014@u.northwestern.edu)

My research interests are in the fields of machine learning, deep learning, and data mining. Particularly, I am interested in large-scale deep learning and its applications. Recently, a myriad of applications adopt deep learning to solve their classification/regression problems. Despite the popularity, there are still many questions remaining unanswered. One crucial question is how to design deep learning methods in a principled way. Most of the deep learning methods for domain problems are designed by 'tuning' many hyper-parameters in a trial-and-error manner. This approach can consume a large amount of time and resources making it less practical. In addition, since users typically tune the hyper-parameters focusing only on the model accuracy, the model tends to be computationally inefficient and not scalable.

My research goal is to tackle large-scale real-world problems by designing efficient and effective deep learning methods. During my Ph.D. study, I was lucky enough to work with many scientists from various fields. I have participated in several deep learning-based scientific projects such as Climate image data restoration, Cosmology data regression, and High-Energy Physics data analysis. In these works, I found that the efficient network training is critical to effectively address the large-scale scientific problems. Motivated by this insight, I study deep learning with a focus on how to pursue both effectiveness and efficiency. In my Ph.D. thesis research, I have studied how to better understand the internal behaviors of neural networks and design scalable training algorithms exploiting such information.

Communication-efficient gradient averaging for parallel training – In order to design a deep and large model for large-scale domain problems, efficient training method is critical to explore many possible design options and find the best model design. In synchronous Stochastic Gradient Descent (SGD) with data parallelism, the most popular parallel training strategy for deep learning, the communications for averaging gradients among all workers are the performance bottleneck. To address this performance issue, I designed a communication-efficient gradient averaging algorithm for data-parallel neural network training [1, 2]. The proposed algorithm relocates the intermediate data, such as activations and errors, across workers so that each worker directly computes the global gradients of a distinct subset of parameters. This approach has a cheaper communication cost than *allreduce*-based approach that is the most popular communication algorithm for gradient averaging. I applied this algorithm to several large-scale scientific applications and successfully scaled up the neural network training on High-Performance Computing platforms.

Adaptive hyper-parameter tuning for scalable deep learning – Mini-batch SGD is the workhorse for deep learning. Most of deep learning applications train their models using the classical mini-batch SGD or the variant algorithms such as Adam [3]. In SGD-based algorithms, however, the degree of parallelism is limited by the mini-batch size. Increasing the batch size improves the degree of parallelism while it can adversely affects the model accuracy. I proposed an adaptive batch size adjustment method to improve the degree of parallelism without a significant accuracy loss [4]. The proposed method monitors the quality of the model with respect to the estimated generalization performance. Then, it gradually increases the batch size when the quality is sufficiently improved at run-time. This approach makes a good trade-off between the model accuracy and the scalability by adaptively adjusting the batch size in the early training epochs. In a collaboration with scientists in Meteorology, I successfully developed a scalable deep learning-based solution to Climate image data restoration problem using the proposed method.

Future work – Based on my graduate research, I plan to further study how to design efficient and effective learning methods for domain problems. Most of the deep learning methods are usually designed based on statistical properties and then applied to specific domain problems. One of the key insights from my experience on several domain projects is that every domain data has its own characteristics such as sparsity, variance across samples, or any inherent data patterns. So, in order to effectively solve such domain problems, I believe that the domain-specific data characteristics should be well studied first and then the learning method should be designed based on such understanding. In addition, mapping the inherent data representation and the internal behaviors of neural networks can improve the interpretability of deep learning methods. Consequently, deep learning methods can be designed in a more principled way rather than 'tuned' in a trial-and-error manner.

Another interesting research topic is how to estimate the generalization performance of neural networks. Most of the training algorithms are designed and analyzed with a focus on their convergence rate. Many researchers have put much effort into designing training algorithms that rapidly minimize the cost function. However, the fast convergence of training loss does not guarantee a good generalization performance. For instance, the variance reduced training algorithms present a significantly improved convergence rate while losing the validation accuracy [5]. If the generalization performance can be precisely estimated during training, the model can be appropriately adjusted at run-time. This approach can considerably improve the efficiency of the model design workflow.

Finally, I am interested in applying efficient and scalable deep learning methods to important real-world problems, such as Computer Vision, Natural Language Processing, and many scientific applications. Considering the ever-increasing available data in this 'bigdata' era, faster training can be considered as a chance to build more powerful models. This research direction includes the topics such as low-precision neural networks [6, 7], gradient compression [8, 9], and scalable parallel training [1, 2, 4]. I believe that developing efficient deep learning solutions will provide researchers with unprecedented opportunities for scientific discoveries in a myriad of domain fields.

References

[1] S. Lee, D. Jha, A. Agrawal, A. Choudhary, and W.-k. Liao, "Parallel deep convolutional neural network training by exploiting the overlapping of computation and communication," in *2017 IEEE 24th International Conference on High Performance Computing (HiPC)*, pp. 183–192, IEEE, 2017.

[2] S. Lee, A. Agrawal, P. Balaprakash, A. Choudhary, and W.-K. Liao, "Communication-efficient parallelization strategy for deep convolutional neural network training," in *2018 IEEE/ACM Machine Learning in HPC Environments (MLHPC)*, pp. 47–56, IEEE, 2018.

[3] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[4] S. Lee, Q. Kang, S. Madireddy, P. Balaprakash, A. Agrawal, A. Choudhary, R. Archibald, and W.-k. Liao, "Improving scalability of parallel cnn training by adjusting mini-batch size at run-time," in *2019 IEEE International Conference on Big Data (Big Data)*, pp. 830–839, IEEE, 2019.

[5] A. Defazio and L. Bottou, "On the ineffectiveness of variance reduced optimization for deep learning. 2018," *Preprint.*

[6] X. Jia, S. Song, W. He, Y. Wang, H. Rong, F. Zhou, L. Xie, Z. Guo, Y. Yang, L. Yu, *et al.*, "Highly scalable deep learning training system with mixed-precision: Training imagenet in four minutes," *arXiv preprint arXiv:1807.11205*, 2018.

[7] Z. Cai, X. He, J. Sun, and N. Vasconcelos, “Deep learning with low precision by half-wave gaussian quantization,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5918–5926, 2017.

[8] D. Miyashita, E. H. Lee, and B. Murmann, “Convolutional neural networks using logarithmic data representation,” *arXiv preprint arXiv:1603.01025*, 2016.

[9] S. Han, H. Mao, and W. J. Dally, “Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding,” *arXiv preprint arXiv:1510.00149*, 2015.