INVESTIGATION ON PARALLELIZATION OF DEEP NEURAL NETWORK TRAINING USING MULTIPLE GPUS

Hang Su[1,2]*, Haoyu Chen*[1]*, Nelson Morgan*[2]

[1] International Computer Science Institute, Berkeley, California, US
[2] Dept. of Electrical Engineering & Computer Science, University of California, Berkeley, CA, USA
{suhang3240@gmail.com, williamchenthu@gmail.com}

ABSTRACT

In this paper we introduce Butterfly mixing to parallel training of deep neural networks (DNN). Parallelization is done in a model averaging manner. Data is partitioned and distributed to different nodes for local model update, and model averaging (reduce) is done every few minibatches. We compare several different reduce stratgies, including all reduce, butterfly mixing, ring reduce and hopping ring reduce. We show that all these methods can effectively speed up neural network training. On swithboard data, a xx times speed up is achieved using 16 gpus.

***Index Terms*—** Parallel training, multiple GPUs, neural network, butterfly mixing

1. INTRODUCTION

Deep Neural Networks (DNN) has shown its effeciveness in several machine learning tasks, espencially in speech recognition. The large model size and massive training examples make DNN a powerful model for classification. However, these two factors also slow down the training procedure of DNNs.

Parallelization of DNN training has been a popular topic since the revive of neural networks. Several different strategies have been proposed to tackle this problem. Multiple thread CPU parallelization and single CUDA GPU implementation are compared in [1, 2], and they show that single GPU could beat 8 cores CPU by a factor of 2.

Optimality for parallelization of DNN training was analyzed in [3], and based on the analysis, a gradient quantization approach was proposed to minimize communication cost [4].

DistBelief proposed in [5] reports a speed up of 2.2x using 8 CPU coresthan using a single machine.

Asynchronous SGD using multiple GPUs achieved a 3.2x speed-up on 4 GPUs [6].

A pipeline training approach was propoased in [7] and a 3.3x speedup was achieved using 4 GPUs, but this method does not scale beyond number of layers in the neural network.

A speedup of 6x to 14x was achieved using 16 GPUs on training convolutional neural networks [8]. In this approach, each GPU is responsible for a partition of the neural network. This approach is more useful for image classification where local structure of the neural network could be exploited.

Distributed model averaging using CPUs is proposed in [9], and a further improvement is done using natural gradient [10].

Butterfly mixing was proposed in [11] to interleave communication with computation.

2. DATA PARALLELIZATION AND MODEL AVERGING

3. DIFFERENT REDUCE STRATEGIES

3.1. Butterfly Mixing

All printed material, including text, illustrations, and charts, must be kept within a print area of 7 inches (178 mm) wide by 9 inches (229 mm) high. Do not write or print anything outside the print area. The top margin must be 1 inch (25 mm), except for the title page, and the left margin must be 0.75 inch (19 mm). All *text* must be in a two-column format. Columns are to be 3.39 inches (86 mm) wide, with a 0.24 inch (6 mm) space between them. Text must be fully justified.

3.2. Ring & Hopping Ring Reduce

The paper title (on the first page) should begin 1.38 inches (35 mm) from the top edge of the page, centered, completely capitalized, and in Times 14-point, boldface type. The authors' name(s) and affiliation(s) appear below the title in capital and lower case letters. Papers with multiple authors and affiliations may require two or more lines for this information. Please note that papers should not be submitted blind; include the authors' names on the PDF.

4. EXPERIMENTAL RESULTS

4.1. Reduce Frequency

4.2. Reduce type

4.3. Scaling factor

5. CONCLUSION

6. ACKNOWLEDGEMENTS

We would like to thank Forrest Iandola for helpful suggestion.

7. REFERENCES

[1] Stefano Scanzio, Sandro Cumani, Roberto Gemello, Franco Mana, and Pietro Laface, "Parallel implementation of artificial neural network training," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 4902–4905.

[2] Karel Veselỳ, Lukáš Burget, and František Grézl, "Parallel training of neural networks for speech recognition," in *Text, Speech and Dialogue*. Springer, 2010, pp. 439–446.

[3] Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu, "On parallelizability of stochastic gradient descent for speech dnns," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 235–239.

[4] Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu, "1-bit stochastic gradient descent and its application to data-parallel distributed training of speech dnns," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[5] Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Mao, Andrew Senior, Paul Tucker, Ke Yang, Quoc V Le, et al., "Large scale distributed deep networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1223–1231.

[6] Shanshan Zhang, Ce Zhang, Zhao You, Rong Zheng, and Bo Xu, "Asynchronous stochastic gradient descent for dnn training," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 6660–6663.

[7] Xie Chen, Adam Eversole, Gang Li, Dong Yu, and Frank Seide, "Pipelined back-propagation for context-dependent deep neural networks.," in *INTERSPEECH*, 2012.

[8] Adam Coates, Brody Huval, Tao Wang, David Wu, Bryan Catanzaro, and Ng Andrew, "Deep learning with cots hpc systems," in *Proceedings of The 30th International Conference on Machine Learning*, 2013, pp. 1337–1345.

[9] Xiaohui Zhang, Jan Trmal, Daniel Povey, and Sanjeev Khudanpur, "Improving deep neural network acoustic models using generalized maxout networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 215–219.

[10] Daniel Povey, Xiaohui Zhang, and Sanjeev Khudanpur, "Parallel training of deep neural networks with natural gradient and parameter averaging," *arXiv preprint arXiv:1410.7455*, 2014.

[11] Huasha Zhao and John Canny, "Butterfly mixing: Accelerating incremental-update algorithms on clusters," in *SIAM Conf. on Data Mining*. SIAM, 2013.