Distributed Deep Learning for Answer Selection

Minwei Feng, Bing Xiang, Bowen Zhou
IBM Watson
Yorktown Heights, NY, USA, 10598
<mfeng|bingxia|zhou>@us.ibm.com

Abstract

This paper is an empirical study of the distributed deep learning for a question answering subtask: answer selection. Comparison studies of SGD, MSGD, DOWNPOUR and EASGD/EAMSGD algorithms have been presented. Experimental results show that the message passing interface based distributed framework can accelerate the convergence speed at a sublinear scale. This paper demonstrates the importance of distributed training: with 120 workers, an 83x speedup is achievable and running time is decreased from 107.9 hours to 1.3 hours, which will benefit the productivity significantly.

1 Introduction

Deep Learning technology [6] has been widely adopted in various AI tasks and has achieved the state-of-the-art performance. One practical challenge of Deep Learning is the highly time consuming training procedure. It is not unusual to see the reported training time in the magnitude of days or even weeks in research papers. However, this is rarely acceptable for practical commercial usage (e.g. training as a service on the cloud) where quick response time is expected by customers. Even for research environment the long time computation could stop scientists from running as many experiments as needed and slow down the R&D cycle. Hence the distributed deep learning training has become a crucial research direction along with the advancement of deep learning itself on the algorithm side.

Various infrastructures and experimental results have been published recently and many of those systems are from leading IT companies. Most of those papers report results on computer vision benchmark tasks like CIFAR10 or ImageNet. In this paper, we focus on the answer selection benchmark task from question answering (QA). It is common to observe the epoch speed (training data processing speed) increased when more computing resources are adopted. However, this does not necessarily guarantee that the convergence speed is also improved. The ultimate goal is the convergence speedup as the customers will expect models with the same accuracy available within much shorter time when the cost is increased for more computing resources. The goal of this paper is to demonstrate the sublinear scalability of the distributed training on convergence speed. We have compared the latest technologies, including SGD [1], MSGD [9], DOWNPOUR [3] and EASGD/EAMSGD [11].

The rest of the paper is organized as follows: section 2 is the summary of related work; section 3 will describe the answer selection benchmark task; we demonstrate the MPI-based infrastructure in section 4 and the review of the distributed training algorithms is given in section 5. Experimental results are reported in section 6 and finally conclusions are drawn in section 7.

2 Related Works

Various systems have been proposed for distributed deep learning. One of the pioneering work is Google’s Distbelief system [3] in which DOWNPOUR has been proposed. The system has multiple parameter servers and clients. Most of other work follow the same spirit of DOWNPOUR. The system Adam [2] is another similar framework which has many engineering features like reduced memory copies and mitigating the impact of slow machines. IBM’s Rudra system [5] is a
Figure 1: Architecture. HL is the hidden layer with \( \text{Tanh} \) activation function. CNN is the convolutional neural networks. P stands for maxpooling and R stands for the \( \text{ReLU} \) activation function. QA means the weights of corresponding layer are shared by Q and A.

Client 1

Client 2

Client n

Server 1

Server 2

Server n

Figure 2: MPI framework

3 Answer Selection Task

This section describes the benchmark test used in this paper. Different from many previous work, we work on a QA task: answer selection. The paper [4] created an open task (including the released corpus) which serves as a benchmark for comparison purpose. For the detailed description of the data and task please refer to [4]. A summary is given here to make the paper self-contained.

Answer selection is an important subtask of QA. Given a question \( q \) and an answer candidate pool \( \{a_1, a_2, ..., a_s\} \) for that question \( s \) is the pool size, the goal is to find the best answer candidate \( a_k, 1 \leq k \leq s \). If the selected answer \( a_k \) is inside the ground truth set (one question could have more than one correct answer) of question \( q \), the question \( q \) is considered to be answered correctly, otherwise it is answered incorrectly. Figure 1 depicts a deep learning approach for this task. The idea is to learn a distributed vector representation of a given question and its answer candidates and then use a similarity metric to measure the matching degree. The similarity metric is Geometric mean of Euclidean and Sigmoid Dot product (GESD) \( k(x, y) = \frac{1}{1+\|x-y\|} \cdot \frac{1}{1+\exp(-|x|y+1)} \). \( x \) and \( y \) are the vector representations of Q and A. The answer candidate which gives the highest similarity value will be selected as the answer output. The training of this approach is computational expensive due to the usage of the hinge loss. During training, for each question \( Q \) there is a positive answer \( A^+ \) (the ground truth). A training instance is then constructed by pairing this \( A^+ \) with a negative answer \( A^- \) (a wrong answer) sampled from the whole answer space. The deep learning framework generates vector representations for the question and the two candidates: \( V_Q, V_{A^+} \) and \( V_{A^-} \). The GESD similarities \( GESD(V_Q, V_{A^+}) \) and \( GESD(V_Q, V_{A^-}) \) are calculated and their difference is compared to a margin \( m \): \( GESD(V_Q, V_{A^+}) - GESD(V_Q, V_{A^-}) < m \). If this condition is not satisfied, there is no update to the model and a new negative example is sampled until the margin is less than \( m \) (this repetitive negative sampling procedure is time-consuming and to reduce running time we set maximum 100 times in this paper).

4 MPI-based Framework

We have chosen Message Passing Interface (MPI) technology to make the master and client communicate with each other. Message passing is designed for application like deep learning where frequent communications need to be conducted and has been reported to be more efficient than other higher level toolkit like Spark [8]. Figure 2 demonstrates our MPI framework. There are two types of process: client and server which are allocated across the high performance computing clusters. The clients conduct the forward pass/backward pass of the neural network and send the update messages to servers. The servers hold a central model. They receive the messages from clients and update the central model and send the latest model back to the clients. In this paper we always

\(1\)https://github.com/dmlc/mxnet
use the non-blocking communication (MPI_ISend/MPI_IRecv) to increase the overall speed. To reduce the communication overhead, we split the model into partitions and set up multiple servers. Each server is responsible for the storage and update of one model partition. In this paper, we take a symmetric structure: the amount of client and server is same. For the method comparison purpose, the accuracy of test corpus should be calculated during the training procedure. Therefore we add a tester which will get the latest model from the servers and run testing over the test corpus periodically. We use the popular toolkit MPICH\footnote{http://www.mpich.org/} in this work.

5 Distributed Training Algorithms

We have compared state-of-the-art algorithms: stochastic gradient descent (SGD) \cite{1}, momentum stochastic gradient descent (MSGD) \cite{2}, DOWNPOUR \cite{3}, elastic averaging stochastic gradient descent (EASGD) \cite{4} and its variation momentum EASGD (EAMSGD) \cite{5}. The DOWNPOUR and EASGD algorithms have been given in Table\ref{table1} and Table\ref{table2}. In both tables the loop is endless because the purpose of this work is to compare the convergence speed. In practice Until forever should be Until stop criteria satisfied. \(\tilde{x}\) represents the central model maintained by the servers. \(x^t\) is the model hold by the client \(i\). \(g^t_i(x^t)\) is the local gradient calculation by client \(i\) at time step \(t^i\). \(\eta\) represents the learning rate. Notice there is a \(\tau\) in both algorithms. It is a crucial hyper parameter which controls the communication frequency. For DOWNPOUR \(\tau\) controls the gradient exchange frequency and for EASGD \(\tau\) controls the elastic difference \(\alpha(x^t - \tilde{x})\) exchange frequency.

| Algorithm 1: DOWNPOUR | Algorithm 2: EASGD |
|-----------------------|---------------------|
| **Input:** learning rate \(\eta\), communication period \(\tau \in \mathbb{N}\) | **Input:** learning rate \(\eta\), moving rate \(\alpha\), communication period \(\tau \in \mathbb{N}\) |
| **Initialize:** randomly initialized \(\tilde{x}\), \(x^t = \tilde{x}\), \(v^t = 0\), \(t^i = 0\) | **Initialize:** randomly initialized \(\tilde{x}\), \(x^t = \tilde{x}\), \(t^i = 0\) |
| **Repeat** | **Repeat** |
| if \((\tau\text{ divides } t^i)\) then | if \((\tau\text{ divides } t^i)\) then |
| \(\tilde{x} \leftarrow \tilde{x} + v^t\) | \(x^t \leftarrow x^t - \alpha(x^t - \tilde{x})\) |
| \(x^t \leftarrow \tilde{x}\) | \(\tilde{x} \leftarrow \tilde{x} + \alpha(x^t - \tilde{x})\) |
| \(v^t \leftarrow 0\) | end |
| \(x^t \leftarrow x^t - \eta g^t_i(x^t)\) | \(x^t \leftarrow x^t - \eta g^t_i(x^t)\) |
| \(v^t \leftarrow v^t - \eta g^t_i(x^t)\) | \(t^i \leftarrow t^i + 1\) |
| Until forever | **Until forever** |

Table 1: DOWNPOUR  Table 2: EASGD

6 Experimental Results

Table\ref{table3} demonstrates the comparison results of convergence speedup. For the method of SGD and MSGD, we let the training keeps running for 5 days. For other methods which use multiple servers and clients, the running time limit is then set to be 12 hours. This is to save the computing resources so that more experiments can be scheduled. Also in practice it is much less meaningful if the running time is still prohibitive when large amount of computing resources are used. As shown in \cite{4}, the accuracy on the test1 corpus is around 63%. Hence for the comparison purpose, we define the convergence speed as the time needed for the accuracy score of test1 corpus to climb up to 63%. SGD is set as the baseline.

Table\ref{table3}a shows the performance of different methods under various learning rates and it is clear that the learning rate makes a crucial impact to the convergence speed. In Figure\ref{fig1} and Figure\ref{fig2} selected runs in Table\ref{table3}a have been plotted. We then fix the learning rate and tune the communication period \(\tau\) in Table\ref{table3}b. Notice in \ref{table3}b DOWNPOUR with \(lr = 0.05\) achieves 57\% speed up. However, in Figure\ref{fig2} it (yellow solid square) shows that DOWNPOUR with \(lr = 0.05\) will diverge in the end. So for the study in Table\ref{table3}b the learning rate is set to be 0.01 for DOWNPOUR. From the results in Table\ref{table3}b we find both EASGD/EAMSGD will suffer from the increased communication period while DOWNPOUR is immune to the performance loss with increased communication period. This is an important property, especially when the client quantity is large which could cause heavy load
within 12 hours (for MSGD it means it cannot reach 63% within 5 days). Column M represents methods; amount of clients (workers). NA means not available which indicates the accuracy cannot climb up to 63% column T indicates the convergence time in hour unit; column S is the speedup times rounded to integer.

Table 3: Convergence speed comparison. The definition of convergence time is the amount of hours needed for the test corpus to reach 63% accuracy. lr is the learning rate. τ is the communication period. w is the amount of clients (workers). NA means not available which indicates the accuracy cannot climb up to 63% within 12 hours (for MSGD it means it cannot reach 63% within 5 days). Column M represents methods; column T indicates the convergence time in hour unit; column S is the speedup times rounded to integer.

| M         | T   | S   | DESCRIPTION |
|-----------|-----|-----|-------------|
| SGD       | 107.9 | w = 1, lr = 0.01 |
| MSGD      | NA  | NA  | w = 1, lr = 0.01 |
| MSGD      | 87.7 | 1   | w = 1, lr = 0.0001 |
| DOWNPOUR | 3.1  | 35  | w = 48, lr = 0.01, τ = 1 |
| DOWNPOUR | NA  | NA  | w = 48, lr = 0.05, τ = 1 |
| EASGD     | 11.4 | 9   | w = 48, lr = 0.01, τ = 1 |
| EASGD     | 7.8  | 14  | w = 48, lr = 0.05, τ = 1 |
| EAMSGD    | 4.2  | 26  | w = 48, lr = 0.0005, τ = 1 |
| EAMSGD    | 3.5  | 31  | w = 48, lr = 0.001, τ = 1 |
| EAMSGD    | 4.5  | 24  | w = 48, lr = 0.005, τ = 1 |
| EAMSGD    | 5.8  | 19  | w = 48, lr = 0.01, τ = 1 |
| EAMSGD    | NA  | NA  | w = 48, lr = 0.05, τ = 1 |

(a) Learning rate comparison
(b) Communication period comparison

Figure 3: Training loss vs. running time

to the network. It means for DOWNPOUR we can achieve faster convergence with more clients and at the same time by increasing the communication period the network load is controlled without performance loss.

The best speedup of various methods are as follows: (1) DOWNPOUR 39 times/48 workers, 47 times/60 workers and 83 times/120 workers; (2) EASGD 14 times/48 workers; (3) EAMSGD 31 times/48 workers.
7 Conclusions

We have conducted an empirical study for the answer selection task which is a crucial component of QA. We build the framework with MPI. The state-of-the-art algorithms have been compared, including SGD/MSGD, DOWNPOUR and EASGD/EAMSGD. To our best knowledge, it is the first time that the experimental results for distributed deep learning have been reported on a QA task. This work proves the significance of the distributed training. A 83x speedup is achievable with the deployment of 120 workers, which is a huge gain for practical productivity. We realize that due to the lack of a solid mathematical foundation, the distributed deep learning is still a trial-and-error procedure. Our experiences show that the hyper parameter tuning can play a crucial role for the convergence speed. On the other hand, the task itself could change the performance. For example, in [11] the EASGD/EAMSGD work better than DOWNPOUR for the image classification tasks while for this answer selection task DOWNPOUR performs better. From Figure 4, we find DOWNPOUR converges faster than EAMSGD while EAMSGD can keep higher accuracy value during the latter half of the 12 hours. For future work we plan to combine EAMSGD and DOWNPOUR so that both methods could benefit from each other and further improvement could be achieved.
Acknowledgments

The authors thank Sixin Zhang for implementation guidance, helpful discussion and valuable feedback.

References

[1] Léon Bottou. Online learning in neural networks. chapter Online Learning and Stochastic Approximations. Cambridge University Press, New York, NY, USA, 1998.

[2] Trishul Chilimbi, Yutaka Suzue, Johnson Apacible, and Karthik Kalyanaraman. Project adam: Building an efficient and scalable deep learning training system. In Proceedings of the 11th USENIX Symposium on Operating Systems Design and Implementation (OSDI 14), pages 571–582, Broomfield, CO, 2014. USENIX Association.

[3] Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Quoc V. Le, Mark Z. Mao, Marc’Aurelio Ranzato, Andrew W. Senior, Paul A. Tucker, Ke Yang, and Andrew Y. Ng. Large scale distributed deep networks. In Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012., pages 1232–1240, 2012.

[4] Minwei Feng, Bing Xiang, Michael R. Glass, Lidan Wang, and Bowen Zhou. Applying deep learning to answer selection: A study and an open task. In Proceedings of the 2015 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2015), Scottsdale, Arizona, 2015.

[5] S. Gupta, W. Zhang, and J. Milthorpe. Model Accuracy and Runtime Tradeoff in Distributed Deep Learning. ArXiv e-prints, September 2015.

[6] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. Nature, 521:436–444, May 2015.

[7] Mu Li, 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), pages 583–598, Broomfield, CO, October 2014. USENIX Association.

[8] Chieh-Yen Lin, Cheng-Hao Tsai, Ching-Pei Lee, and Chih-Jen Lin. Large-scale logistic regression and linear support vector machines using spark. In 2014 IEEE International Conference on Big Data, pages 519–528, October 2014.

[9] Ilya Sutskever, James Martens, George E. Dahl, and Geoffrey E. Hinton. On the importance of initialization and momentum in deep learning. In Proceedings of the 30th International Conference on Machine Learning (ICML-13), volume 28, pages 1139–1147. JMLR Workshop and Conference Proceedings, May 2013.

[10] Eric P. Xing, Qirong Ho, Wei Dai, Jin-Kyu Kim, Jinliang Wei, Seunghak Lee, Xun Zheng, Pengtao Xie, Abhimanyu Kumar, and Yaoliang Yu. Petuum: A new platform for distributed machine learning on big data. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’15, pages 1335–1344, 2015.

[11] Sixin Zhang, Anna Choromanska, and Yann LeCun. Deep learning with elastic averaging SGD. In Proceedings of the 2015 Conference on Neural Information Processing Systems. (NIPS 2015), 2015.