联邦学习中的通信优化

叶茂青，王珺

June 23, 2020

Abstract

随着机器学习的不断发展，机器学习所需要的数据量也不断增加，但客户对于数据隐私的需要，使得机器学习所需的数据获取难度很高，联邦学习试图在保护用户数据的同时，建立出能满足用户需求的模型。但大规模的联邦学习训练对带宽要求极高，对于个人设备，高昂的通信成本更是限制了复杂模型的使用，因此，提高联邦学习中客户端与服务器端的通信效率极为重要。本文总结了在联邦学习中针对网络梯度传输所做的各种优化。

1 联邦学习的介绍

联邦学习的数学模型被 McMahan et al. [1]，McMahan et al. 同时也提出了联邦学习区别于分布式学习的四大难点：

- Non-IID
- Unbalanced
- Massively distributed
- Limited communication

2 梯度压缩

针对

2.1 Gradient quantization

Gradient quantization 的思路主要是 xxx

2.2 Gradient sparsification

Gradient sparsification 的思路主要是 xxx

	Works
Gradient quantization	Wen et al. [2] Seide et al. [3] Zhou et al. [4]
Gradient sparsification	Storm [5] Dryden et al. [6] Aji & Heafielf [7] Chen et al. [8]

References

[1] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017*, feb 2016. [Online]. Available: http://arxiv.org/abs/1602.05629

[2] W. Wen, C. Xu, F. Yan, C. Wu, Y. Wang, Y. Chen, and H. Li, "TernGrad: Ternary gradients to reduce communication in distributed deep learning," Tech. Rep., 2017. [Online]. Available: https://github.com/wenwei202/terngrad

[3] F. Seide, H. Fu, J. Droppo, G. Li, and D. Yu, "1-bit stochastic gradient descent and its application to data-parallel distributed training of speech DNNs," Tech. Rep., 2014. [Online]. Available: https://www.isca-speech.org/archive/interspeech_2014/i14_1058.html

[4] S. Zhou, Y. Wu, Z. Ni, X. Zhou, H. Wen, and Y. Zou, "DoReFa-Net: Training Low Bitwidth Convolutional Neural Networks with Low Bitwidth Gradients," Tech. Rep., 2016. [Online]. Available: http://arxiv.org/abs/1606.06160

[5] N. S. S. A. C. of the International and undefined 2015, "Scalable distributed DNN training

using commodity GPU cloud computing," *isca-speech.org*. [Online]. Available: https://www.isca-speech.org/archive/interspeech_2015/i15_1488.html

[6] N. Dryden, T. Moon, S. A. Jacobs, and B. Van Essen, "Communication quantization for data-parallel training of deep neural networks," *Proceedings of MLHPC 2016: Machine Learning in HPC Environments - Held in conjunction with SC 2016: The International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 1–8, 2017. [Online]. Available: https://github.com/LLNL/lbann.

[7] A. F. Aji and K. Heafield, "Sparse communication for distributed gradient descent," in *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*. Association for Computational Linguistics (ACL), 2017, pp. 440–445.

[8] C. Y. Chen, J. Choi, D. Brand, A. Agrawal, W. Zhang, and K. Gopalakrishnan, "ADaComP: Adaptive residual gradient compression for data-parallel distributed training," *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pp. 2827–2835, 2018. [Online]. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/viewPaper/16859