联邦学习中的通信优化

17363092 叶茂青，17363079 王珺

June 24, 2020

Abstract

联邦学习的提出使得用户可以在保证隐私的同时贡献自己的数据，但高昂的通信成本限制了联邦学习的应用，本文总结了联邦学习针对网络梯度传输所做的各种优化。

1 TODO

problem that the research addresses
background information and relevant references
elements that validate the level of innovation of the research
conceptual model, methodology or procedure that the research takes into consideration
analysis and interpretation of the results achieved
strengths and weaknesses of the research, the insights demonstrated
implications for further research

2 联邦学习的介绍

联邦学习的数学模型描述在 2016 年由 McMahan et al. [1] 提出，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

由于联邦学习的应用场景中，客户端多为个人设备，通信成本高且通信质量也不能保证

3 梯度压缩

针对

	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]

3.1 Gradient quantization

Gradient quantization 的思路主要是将原来 32 位浮点数存储的梯度信息量化为更低准确度的存储表示。这方面的研究有 Seide et al. [3] 提出的 1-bit SGD、Alistarh et al. [9] 提出的 QSGD、Wen et al. [2] 提出的 TernGrad、Zhou et al. [4] 提出的 DoReFa-Net、Storm et al. [5] 提出的 xxx

Gradient quantization 的难点在于如何在压缩梯度后依然能保证网络的收敛，对此 Seide et al. [3] 使用 Error Feedback 的方法，将每次的量化误差保存下来并再次反馈到量化函数中

虽然 1-bit SGD 在实验观察下依旧可以保持收敛，但在其他条件下的收敛性依旧存疑，Alistarh et al. [9] 提出的 QSGD 则用数学证明了其在任意函数下的收敛性，在实验的表现中，使用 QSGD 对梯度进行 4Bit 或 8Bit 量化，无论是在 resnet 还是 lstm 上的精度都和 32BitSGD 没有太大区别，同时实验也提出了卷积层对于量化的敏感性或许要更高。QSGD 允许用户在通信量和训练时间进行权衡

Wen et al. [2] 提出的 TernGrad 将梯度为 2Bit，使用-1,0,1 三个值表示梯度。其数学表示为

$$\tilde{\boldsymbol{g}}_t = \text{ternarize}\left(\boldsymbol{g}_t\right) = s_t \cdot \text{sign}\left(\boldsymbol{g}_t\right) \circ \boldsymbol{b}_t$$

where

$$s_t \triangleq \max\left(abs\left(\boldsymbol{g}_t\right)\right) \triangleq \left\|\boldsymbol{g}_t\right\|_\infty$$

其中 b_t 满足伯努利分布

$$\begin{cases} P\left(b_{tk} = 1 \mid \boldsymbol{g}_t\right) = \left|g_{tk}\right| / s_t \\ P\left(b_{tk} = 0 \mid \boldsymbol{g}_t\right) = 1 - \left|g_{tk}\right| / s_t \end{cases}$$

为了避免某些梯度过大导致绝大部分梯度被量化为 0，作者引入 layer-wise ternarizing，逐层的对梯度进行量化，同时对于不同层之间梯度方差的差异，作者引入 gradient clipping，用方差信息对每一层的梯度进行裁剪。

作者使用在 MNIST 和 CIFAR-10 上训练的卷积网络进行验证，相比 32BitSGD，使用 TernGrad 并没有减慢收敛的速度，且精度损失也在 1% 以内。

3.2 Gradient sparsification

Gradient sparsification 的思路主要是 xxx [] [10]

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

[9] D. Alistarh, D. Grubic, J. Z. Li, R. Tomioka, and M. Vojnovic, "QSGD: Communication-efficient SGD via gradient quantization and encoding," Tech. Rep., 2017. [Online]. Available: http://papers.nips.cc/paper/6768-qsgd-communication-efficient-sgd-via-gradient-quantization-and-encoding

[10] Y. Lin, S. Han, H. Mao, Y. Wang, and W. J. Dally, "Deep Gradient Compression: Reducing the Communication Bandwidth for Distributed Training," dec 2017. [Online]. Available: http://arxiv.org/abs/1712.01887