DEEP GRADIENT COMPRESSION: REDUCING THE COMMUNICATION BANDWIDTH FOR DISTRIBUTED TRAINING

Wang Jian

2020 年 11 月 16 日

目录

1 Author

林宇🄴是清华大学电子工程系 NICS 实验室 2014 级本科生，于 2017 年暑假在斯坦福参加暑研期间同韩松博士一起出色完成了 DGC 的工作，收到 MIT, Stanford, CMU, UMich 等美国名校的博士项目录取，并将于 2018 年秋加入 MIT HAN Lab 攻读博士学位。

韩松博士于 2017 年毕业于斯坦福大学，师从 GPU 之父 Bill Dally 教授。他的研究涉足深度学习和计算机体系结构，他提出的 Deep Compression 模型压缩技术曾获得 ICLR 2016 最佳论文，ESE 稀疏神经网络推理引擎获得 FPGA 2017 最佳论文，引领了世界深度学习加速研究，对业界影响深远，

于博士期间联合创立了深鉴科技。基于对一系列重要科研成果的继续深入探索，韩松博士将于 2018 年任职 MIT 助理教授，创立 HAN Lab。

2 Background

- Large-scale distributed training requires significant communication bandwidth for gradient exchange.
- 99.9% of the gradient exchange in distributed SGD are redundant.
- Preserve accuracy during compression.

3 Main Work

- propose Deep Gradient Compression(DGC) to reduce the communication bandwidth.
- DGC employs 4 methods to preserve accuracy.
  - momentum correction, local gradient clipping, momentum factor masking, and warm-up training.

4 Comments

- 实验结果令人信服