一种基于统计学方法的分布式训练系统的高效梯度压缩技术

AN EFFICIENT STATISTICAL-BASED GRADIENT COMPRESSION TECHNIQUE FOR DISTRIBUTED TRAINING SYSTEMS

Introduction

深度神经网络的不断增大和愈加复杂，使得人们引入了分布式学习系统来处理和训练大型网络，以期望得到有效的模型。由于模型训练的同步性质，在第`i`次梯度聚合可用之前，worker节点无法进行第`i+1`次训练。同时，模型训练的瓶颈也往往受限于通信瓶颈。

本文通过对通讯时的梯度进行有损压缩从而克服通讯瓶颈。

主要挑战在于：

1. 有损压缩带来的梯度影响
2. 引入了额外的计算延迟（此前的诸多工作往往忽略了这一点）

通过以上因素，本文提出了基于空间诱导分布的压缩(SIDCo)，它基于信号可压缩理论，同时支持使用GPU等相关硬件进行并行运算。

Relation Work

以往的工作中，人们通常会努力扩大通讯与计算的重叠度，从而隐藏通讯开销。或者减少通讯的频率，数据量。本文只考虑梯度压缩，因为它能带来很好的收益。

梯度压缩

梯度压缩通过压缩函数对输入梯度进行压缩，其中往往包括*梯度量化*和*梯度空间化*。

梯度量化：使用较少的比特位来表示梯度，但受32位浮点数的限制，在慢速网络和大规模模型上表现不足，需要引入昂贵的量化位。

梯度空间化：选取梯度空间的子集，相关研究证明，一些情况下99.9%的梯度丢失对模型没有影响，Top_k相比于其他类似的压缩算法有更好的收敛效果，但其计算方式或其变体往往会引入庞大的计算代价。

Background and Motivation

Top_k计算优化

为了减少Top_k的计算的代价，人们从算法，硬件上进行优化，但总体上还是对GPU不够友好。

对于Top_k的计算，人们提出了阈值估计算法，优化了计算性能，但算法复杂性没有优化，理论上这种方法能够达到线性复杂度。

此前相关工作对于Top_k的阈值估计算法进行了努力，RedSync(通过移动计算最大梯度和平均梯度的比例计算阈值)，GaussianKSGD(通过拟合高斯分布中获取初始阈值)，但结果偏差较大。

本文中利用统计学方法和选取合适的稀疏诱导分布来准确估计阈值。

Top_k，DGC(随机子样本梯度估计)，RedSync，GaussianKSGD作为SIDCo的对比，显示出良好的加速比。

Contributions

- 提出了一种基于阈值峰值（PoT）的多级拟合技术，该技术适用于积极的稀疏率，并适应梯度的分布变化。
- 设计了SIDCo，这是一种阈值稀疏方法，具有三个SID的闭式表达式，以尽可能降低压缩开销。
- 通过对不同基准的数据集和经验评估，表明SIDCo的表现一直优于现有方法。

Gradient Model and Threshold Estimation

Gradient compressibility

由于信号具有稀疏性同时符合幂律衰减的特点，则这个信号可以被压缩。

在DNN上训练得到的梯度往往是符合幂律衰减的，因此它是可以被压缩的。

Gradient Model

DNN训练过程中产生的梯度可以建模为一些稀疏诱导分布，双指数，双伽马和双广义帕累托分布。

例如在Res-Net-20上进行SGD，将收集到的经验分布用double-exp, double-gamma和double-GP进行拟合，可以得到

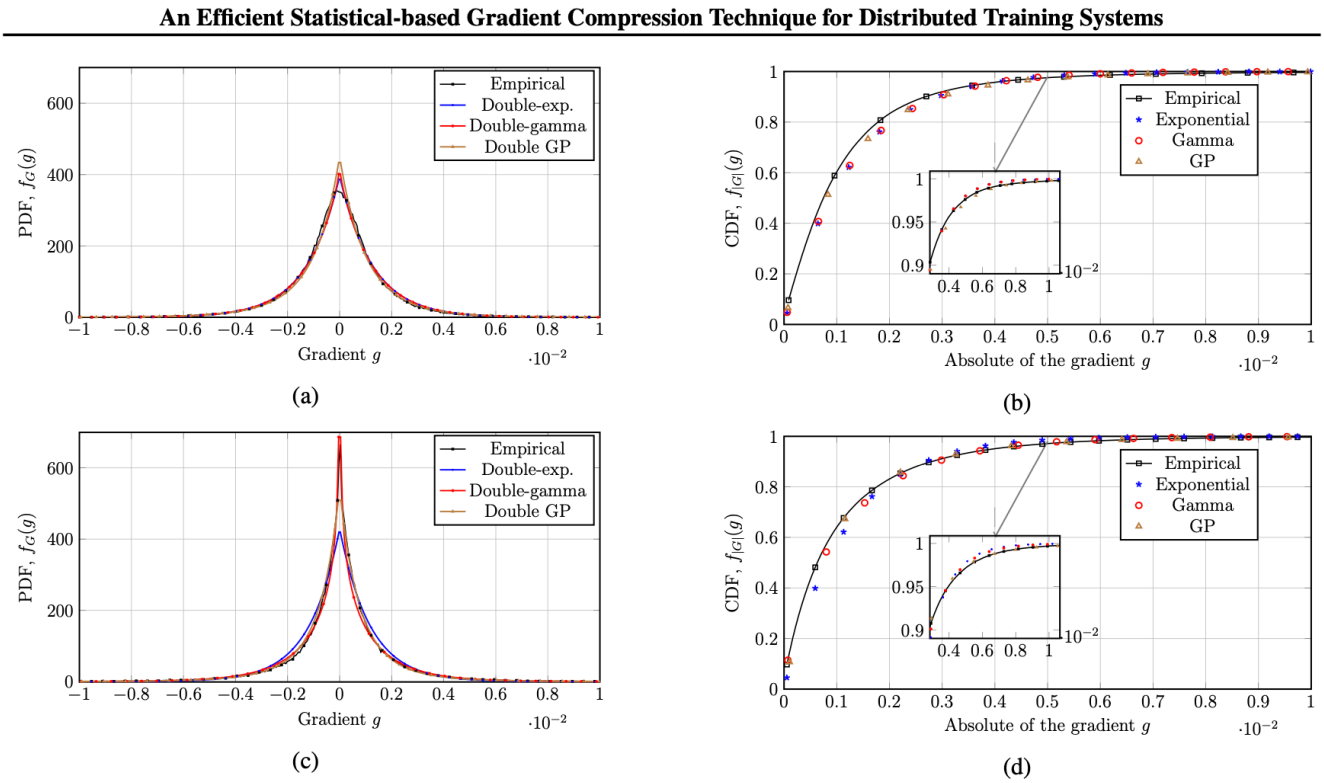


Figure 2. Fitting using the three SIDs for the gradient vector along with the empirical distribution generated from training ResNet-20 on CIFAR10 using Top_k compressor without EC mechanism, for the 100^{th} [(a) PDF, (b) CDF] and 10000^{th} [(c) PDF, (d) CDF] iterations.

迭代10000次的概率分布函数(PDF)比100次的PDF更加稀疏而且拥有更快的尾部。在累积经验分布(CDF)中，它们很好的近似于经验分布，但是在分布尾部，它们倾向于略微高估/低估CDF。

Single-Stage Threshold Estimator

针对3种SID导出它们包含目标压缩比的阈值，在此基础上提出了一种中等压缩比的单级阈值方案。

$$\eta(\delta) = F_{|G|}^{-1}(1-\delta;\widehat{\boldsymbol{\Theta}}) \tag{4}$$

$$= F_G^{-1}\left(1-\frac{\delta}{2};\widehat{\boldsymbol{\Theta}}\right), \tag{5}$$

远端估计可能存在的问题：由于压缩比可能低于10^{-4}，所以为了准确估计阈值，拟合的梯度应与尾部的梯度分布紧密相似，但在一定的压缩比下，单级拟合得到的阈值是准确的。

Multi-Stage Threshold Estimator

通过多阶段你和的方法可以来克服远尾估计问题，首先选择一个SID进行拟合，进行压缩后计算阈值以及压缩比，此后使用超越梯度向量拟合第二个分布，选择第二压缩比使得总体压缩比为目标压缩比。该过程可以推广到多级，即：

$$\delta = \prod_{m=1}^{M} \delta_m$$

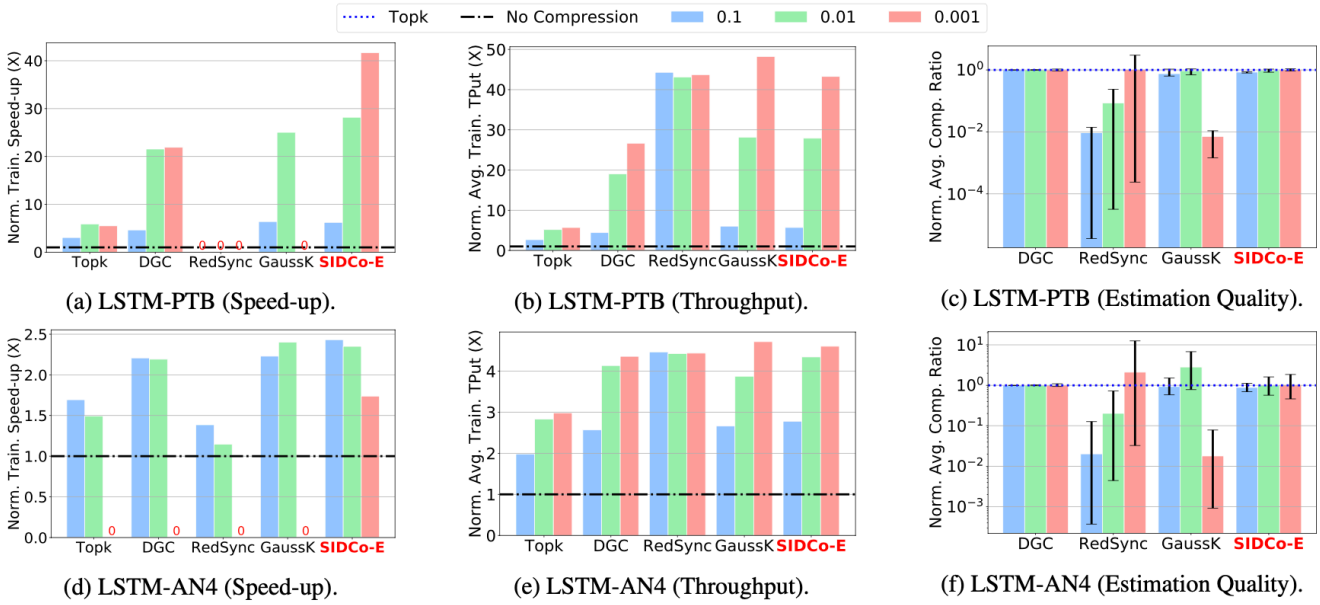
M为级数，δ_m为第m级的压缩比，之后便可利用极值理论来估计多级阈值。

SIDCo Algorithm

通过上文提出的多阶段阈值估计器，可以获得平均误差低于误差边界的阈值估计，具体为：在每次迭代中，将梯度矢量作为输入，产生一个压缩向量。该向量通过上文的多阶段拟合策略进行稀疏。在每个阶段使用所选择的SID进行阈值估计。同时通过估计元素选择的质量，根据误差范围边界，动态的调整级数M。

Experimental Evaluation

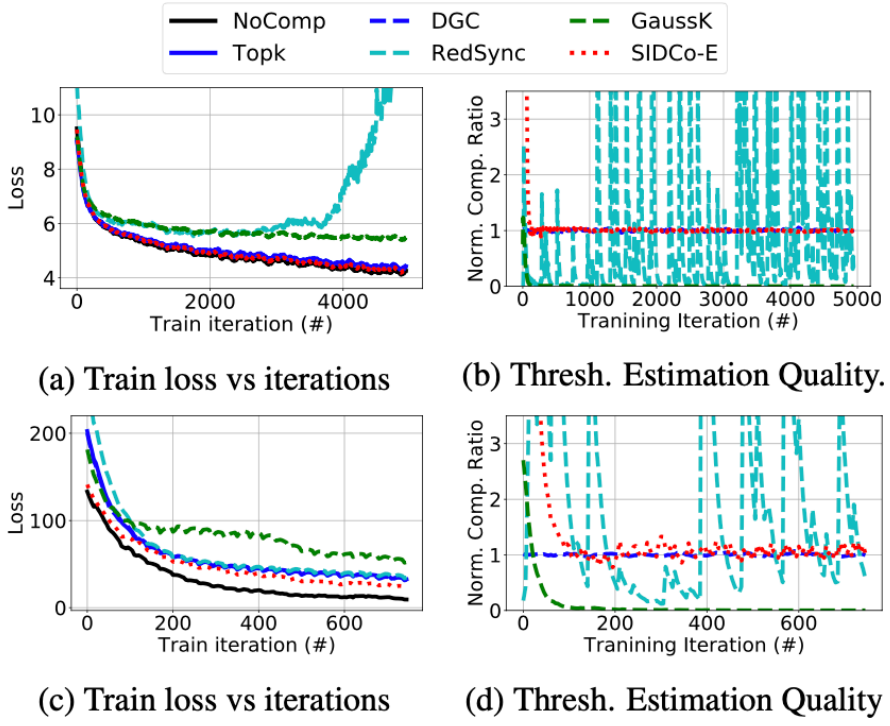
实验分别使用CNN和RNN模型进行图像分类和语言建模任务，其中分别采用压缩比为0.1, 0.01, 0.001进行压缩和精度的权衡。为了简化展示，实验中仅使用了双指数SID来对比实验效果(SIDCo-E)。



(a) LSTM-PTB (Speed-up). (b) LSTM-PTB (Throughput). (c) LSTM-PTB (Estimation Quality).

(d) LSTM-AN4 (Speed-up). (e) LSTM-AN4 (Throughput). (f) LSTM-AN4 (Estimation Quality).

Figure 3. Performance of training RNN-LSTM on PTB [(a),(b),(c)] and AN4 [(d),(e),(f)] datasets.



(a) Train loss vs iterations (b) Thresh. Estimation Quality.

(c) Train loss vs iterations (d) Thresh. Estimation Quality

Figure 4. The training performance for the LSTM model on PTB and AN4 datasets with compression ratio of 0.001.

实验从模型训练的加速比，吞吐量，预估准确率三个方面进行评估，并于基准线进行比较。

PTB模型具有较大的通讯开销，从实验结果中可以发现，SIDCo具有显著的加速比(图3a)，同时在0.001的压缩率下，RedSync和GassKSGD都是不收敛的(图4a)，所以他们的加速比为0。图3b中显示出良好的吞吐量，但是只有DGC和SIDCo能够准确估计目标比率，其他两种方法方差波动较大，不利于收敛。从图4b中可以看到，Redsync在不断波动，且不收敛；GaussKSGD则导致了非常低的压缩比率。

在AN4上，SIDCo具有优于其他方法的加速比，同时在0.001的压缩比率下，只有SIDCo方法在损失边界下达到目标字符错误率。其他分析同PTB模型上的特点相同。

Takeaways

文中所提出的方法简单但富有成效，相比于其他工作，它们要么没有利用梯度的统计特性，要么在没有深入了解梯度的情况下假定为高斯分布。同时在GPU上，SIDCo取得了更多的加速比效果。作为门限估计方法，SIDCo在保证高吞吐率的同时也提供了高质量的门限估计。

Conclusion

SIDCo解决了在较小计算代价的条件下获得良好的梯度压缩效果，并且在部分CNN和RNN模型上取得了良好的效果。对于其他特点的模型还需要进一步探索。

思考

- 全文对已有工作做了深入剖析，点明了已有工作的不足的地方
- 全文展示了严格的数学推导工作，包括梯度分布的先验证明，阈值估计的推导等，同时提供了大量的图表等信息，对实验背景都提供了详细的描述。
- 对于不同的模型，压缩算法/压缩率的选定也不一定是一致的，压缩算法/压缩比率的错误选择可能会导致模型最终难以收敛。这对于模型训练者来说，可能是一件难以完全透明化的工作。