Global QSGD: Practical Floatless Quantization for Distributed Learning with Theoretical Guarantees

Jihao Xin  
KAUST

Peter Richtárik  
KAUST

Marco Canini  
KAUST

Samuel Horváth  
MBZUI

1 Introduction

The increase of deep learning models and dataset sizes make training time-consuming, while heavy communication, primarily due to gradients synchronization as a key bottleneck. Sapio et al. [7] show that communication can take up to 90% of a training iteration. A popular remedy is to reduce the size of communication data between nodes by applying gradient compression methods [1, 3, 6, 8, 9, 11]. Unfortunately, a majority of the proposed compressors are not natively compatible with the AllReduce collective communication operator because of the change in data format and the need for custom reduction operations. To the best of our knowledge, the only compressors compatible with AllReduce are PowerSGD [10] and IntSGD [5, 7]. However, practical implementations of these methods are heuristic-based and do not come with rigorous theoretical guarantees. Concurrently, C-Coll [4] proposes error-bounded lossy compression with MPI collectives. We address this question: can we provide theoretical guarantees for gradient compressors while retaining AllReduce compatibility for an efficient implementation?

2 Global Quantization

Our main contribution is a new compressor, Global Quantization (G-Q), which quantizes 32-bit floats to smaller bit-widths utilizing the norm of the global gradient such that the quantized data is AllReduce-compatible. Assume gradient $x$, we define $G-Q$ as follows.

The global quantization operator with respect to the $p$ norm and $s$ levels

$$0 = l_{s} < l_{s-1} < l_{s-2} < \cdots < l_{1} < l_{0} = 1,$$

denoted $G-Q_{l_{j}}^{p}$, is defined as follows. Let $x = [x_{1}, x_{2}, \ldots, x_{n}] \in \mathbb{R}^{nd}$. Let $y_{i} \overset{\text{def}}{=} \|x_{i}\|_{q,p} \in \mathbb{R}^{d}$ for all $i \in [n]$. Then

$$G-Q_{l_{j}}^{p}(x) \overset{\text{def}}{=} \|x\|_{q,p} \frac{1}{n} \sum_{i=1}^{n} \text{sign}(x_{i}) \circ \xi_{i}(y_{i}),$$

(1)

where $\xi_{i}(y_{i})$ is an independent element-wise random rounding operator such that

$$(\xi_{i}(y_{i}))_{j} \overset{\text{def}}{=} \begin{cases} l_{u_{j}^{i}} & \text{with probability } \frac{(u_{j}^{i})_{j} - l_{u_{j}^{i}+1}}{u_{j}^{i} - l_{u_{j}^{i}+1}}, \\ l_{u_{j}^{i}+1} & \text{otherwise} \end{cases},$$

(2)

for $j \in [d]$, where $u_{j}^{i} \in \{0, 1, 2, \ldots, s\}$ is such that $l_{u_{j}^{i}} \leq (y_{i})_{j} \leq l_{u_{j}^{i}+1}$.

We adopt two ways to cut the quantization intervals $l$:

- **Standard Dithering**, with linear levels, i.e., $l_{i} = s^{-i}/s$.
- **Exponential Dithering** [3], with exponential levels, i.e., $l_{i} = \psi/2^{s-i}$ for $i \in \{1, \ldots, s\}$.\(^{1}\)

Standard dithering is intrinsically linearly divided, and so can be aggregated directly using a fixed-point reduction operation in existing AllReduce implementations. We also propose a custom reduction operation for exponential dithering, which does not require dequantization. Thus, compressed data can be reduced through AllReduce primitive with the compression ratio up to $O(\sqrt{nd})$, where $n$ is the number of computing nodes and $d$ is the data size. On the other end, traditional quantization methods have to utilize AllGather whose compression ratio can only reach $O(\sqrt{d})$. We theoretically prove that global quantization is unbiased (i.e., error feedback is not required) and with bounded variance. We show that existing works with an unbiased compressor with bounded variance can seamlessly extend to global quantization with the same rate of convergence.

3 Evaluation

We implement $G-Q$ as a drop-in module for PyTorch DDP through a Python extension with CUDA offloads. We run experiments in a server with 4 A100 GPUs communicating via both NVLink and PCIe; we observe that our algorithm is beneficial even with the extremely fast NVLink. We also run large-scale validation in Google Cloud Platform (GCP) with 64 servers each equipped with 1 A100 GPU. Figure 1 shows that global quantization reaches $3.16 \times$ training speedup without loss of accuracy for the DeepLight model [2].

\(^{1}\)We can work with any basis. We use base 2 for simplicity and the fact that this is naturally compatible with the binary representation of floats.
References


[8] Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu. 2014. 1-Bit Stochastic Gradient Descent and Application to Data-Parallel Distributed Training of Speech DNNs. In INTERSPEECH.

