Distributed Stochastic Gradient Descent with Quantized Compressive Sensing

Dipayan Mitra, Ashish Khisti

University of Toronto

Overview

- Millions of connected devices generating huge amount of unprocessed data.
- Distributed processing is adopted for training large scale machine learning models.
- Sync-SGD is a preferred optimization technique [4].
- Gradient communication between the parameter server and the worker causing bottleneck.

Way out: **Gradient compression**

Motivations

- Sparsity induced by ReLU activation function.

  $f(x) = \max(0, x) \quad \rightarrow$ Forcing all $x < 0$ to 0
- 44% of operations performed in most of the modern DNNs, for example AlexNet, GoogLeNet etc., are ineffective [3].

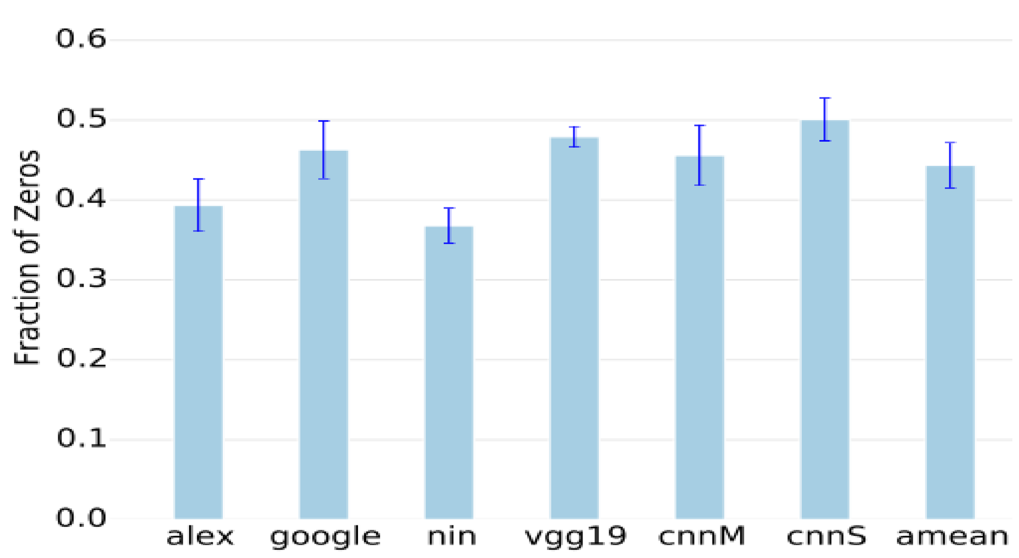

Fig. 1: Average fraction of zero input neuron values in convolutional layer multiplication [3]

Key Idea: **Use quantized compressive sensing to exploit the sparsity.**

Compressive Sensing

- A sampling technique for signals which are sparse or compressible in some known basis [2].
- Measurement matrix $\Phi_{M\times N}$ is chosen to be a random matrix to obtain measurement vector $\mathbf{y}_{M\times 1}$ from signal $\mathbf{x}_{N\times 1}$ as,

$$\mathbf{y}_{M\times 1} = \mathbf{\Phi}_{M\times N}\mathbf{x}_{N\times 1}$$

- Signal is recovered by solving LP optimization problem as follows:

$$\widehat{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmin}} \|\mathbf{x}\|_1 \text{ s.t. } \mathbf{y} = \Phi\mathbf{x}$$

Quantized Compressive Sensing

- Quantization is modelled as as an additive measurement noise in quantized compressive sensing [1]: $\mathbf{y} = Q(\mathbf{\Phi x}) = \mathbf{\Phi x} + \mathbf{e}$
- Measurement noise $\mathbf{n}$ is bounded by the quantization interval Δ and the dimension of the compressed measurement (M): $||\mathbf{e}||_2 \le \sqrt{\frac{M\Delta^2}{12}} = \epsilon$
- Signal reconstructed by solving an optimization problem.

$$\widehat{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmin}} \|\mathbf{x}\|_1 \text{ s.t. } \|\mathbf{y} - \Phi\mathbf{x}\|_2 \le \epsilon$$

- Reconstruction error $||\hat{\mathbf{x}} - \mathbf{x}||_2 = ||\mathbf{n}||_2 \le \beta$

Vanilla Sync-SGD

- K workers participating in a distributed learning to evaluate parameters $\mathbf{w}$.
- Each worker computes its local gradients $\mathbf{g}_t^{(k)}$ and sends to the parameter server to perform aggregation:

$$\mathbf{g}_t = \frac{1}{K}\sum_{k=1}^{K} \mathbf{g}_t^{(k)}$$

- Model parameters are updated following: $\mathbf{w}_{t+1} = \mathbf{w}_t - \gamma\mathbf{g}_t$ and sent back to the workers.

Proposed Approach

- Use quantized compressive sensing to compress the sparse gradients.
- Quantized compressed measurement vectors $\mathbf{y}_t^{(k)}$ are obtained for each worker (k).

$$\mathbf{y}_t^{(k)} = Q(\mathbf{\Phi g}_t^{(k)})$$

- Quantized compressed measurements $\mathbf{y}_t^{(k)}$ are sent to the parameter server.
- At the parameter server the quantized compressed measurements are recovered to obtain $\tilde{\mathbf{g}}_t^{(k)}$.
- Parameter server performs gradient aggregation: $\tilde{\mathbf{g}}_t = \frac{1}{K}\sum_{k=1}^{K} \tilde{\mathbf{g}}_t^{(k)}$.
- Parameters are updated following the update rule and sent back to each worker.

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \gamma\tilde{\mathbf{g}}_t$$

Advantage: Quantization is performed on the compressed gradients lowering communication cost over standard gradient quantization approaches (where quantization is performed directly on the gradients).

Convergence Analysis

- **(Lower bound assumption)** $\forall \mathbf{w}$ and some constant f^*, global objective function $f(\mathbf{w}) > f^*$.
- **(Smoothness assumption)** Let $\bar{\mathbf{g}}(\mathbf{w})$ denote $\nabla \mathbf{f}(\mathbf{w})$ evaluated at $\mathbf{w} = [w_1, w_2, \dots, w_d]^T$. Then $\forall \mathbf{w}$, $\mathbf{\Theta} = [\theta_1, \theta_2, \dots, \theta_d]^T$ and a non-negative constant vector $\mathbf{L} = [l_1, l_2, \dots, l_d]^T$ and $l' = ||\mathbf{L}||_\infty$,

$$|f(\Theta) - [f(\mathbf{w}) + \bar{\mathbf{g}}(\mathbf{w})^T(\Theta - \mathbf{w})]| \le \frac{1}{2}\sum_{i=1}^{d} l_i(\theta_i - w_i)^2$$

- **(Variance bound assumption)** Stochastic gradient $\mathbf{g}(\mathbf{w})$ is an unbiased estimate having bounded coordinate variance $\mathbb{E}[\mathbf{g}(\mathbf{w})] = \bar{\mathbf{g}}(\mathbf{w})$ and,

$$\mathbb{E}[(\mathbf{g}^{(\mathbf{k})}(\mathbf{w})_i - \bar{\mathbf{g}}(\mathbf{w})_i)^2] \le \sigma_i^2$$

  for some non-negative constant vector $\boldsymbol{\sigma} = [\sigma_1, \sigma_2, \dots, \sigma_d]^T$.
- Let $\bar{\mathbf{n}}_t = \mathbb{E}[\mathbf{n}_t]$ and there exists a non-negative μ such that $(\mu < 1)$,

$$||\bar{\mathbf{n}}_t|| \le \mu||\bar{\mathbf{g}}_t||$$

Theorem 1. Let T be the total number of iterations and learning rate $\gamma = \frac{1}{l'K\sqrt{T}}$ and f_0 be the initial objective value. Then,

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} ||\bar{\mathbf{g}}_t||^2\right] \le \frac{1}{\sqrt{T}}\left[\frac{l'K^2(f_0 - f^*) + ||\boldsymbol{\sigma}||^2 + \beta}{1 - \mu}\right]$$

Comparison

- SGD has same asymptotic convergence rate of $\mathrm{O}\left(\frac{\beta}{\sqrt{T}}\right)$ as of our approach.
- TernGRAD provided probabilistic guarantee on convergence [5].
- Error compensated DoubleSqueeze admits the same asymptotic convergence rate of $\mathrm{O}\left(\frac{\beta}{\sqrt{T}}\right)$.

References

[1] Petros Boufounos and Richard Baraniuk. “1-Bit compressive sensing”. In: *42nd Annual Conference on Information Sciences and Systems.* 2008, pp. 16–21.

[2] D Donoho. “Compressed sensing”. In: *IEEE Transactions on Information Theory* 52.4 (2006), pp. 1289–1306.

[3] Albericio *et al.* “Cnvlutin: Ineffectual-Neuron-Free Deep Neural Network Computing”. In: *IEEE ISCA.* 2016, pp. 1–13.

[4] Chen *et al.* “Revisiting Distributed Synchronous SGD”. In: *ArXiv* abs/1604.00981 (2017).

[5] Wei Wen et al. “TernGrad: Ternary Gradients to Reduce Communication in Distributed Deep Learning”. In: *NIPS.* 2017.