DISTRIBUTED STOCHASTIC GRADIENT DESCENT WITH QUANTIZED COMPRESSIVE SENSING

Dipayan Mitra, Ashish Khisti

Department of Electrical and Computer Engineering, University of Toronto

Outlines

INTRODUCTION

1. Millions of connected devices generating huge amount of unprocessed data.
2. How to train a large enough machine learning model without centralizing data?

$\rightarrow$ Distributed processing exploiting data-parallelism.

3. Distributed processing is adopted for training large scale machine learning models.
4. How to optimize such a model?

$\rightarrow$ Synchronous SGD (Sync-SGD).

Sync-SGD

Until Convergence:

Server:

1. Select K number of users randomly.
2. Send $\mathbf{w}_t$, i.e. parameter update at t^{th} iteration, to all K users.

   User:

   1. Download parameter update $\mathbf{w}_t$ from the server.
   2. Run SGD locally (on the local dataset) and obtain $\mathbf{g}_t^{(k)}$.
   3. Upload $\mathbf{g}_t^{(k)}$ to the server.
3. Aggregate gradients: $\mathbf{g}_t = \frac{1}{K}\sum_{k=1}^{K}\mathbf{g}_t^{(k)}$.
4. Update model parameters: $\mathbf{w}_{t+1} = \mathbf{w}_t - \gamma\mathbf{g}_t$

Sync-SGD

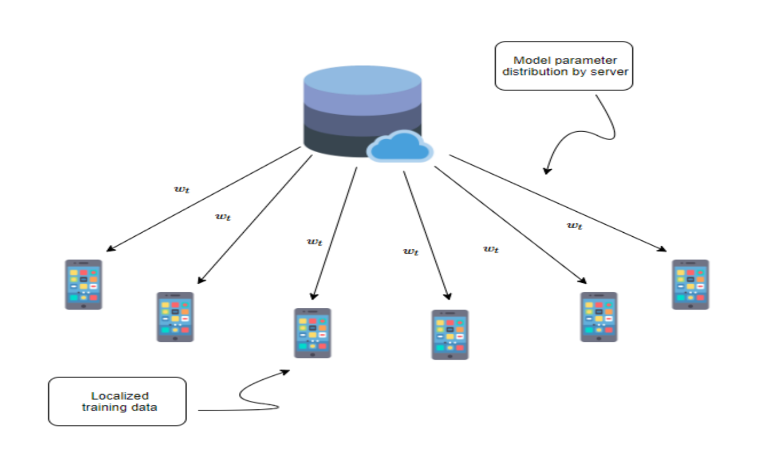

Figure: Sync-SGD (parameter download)

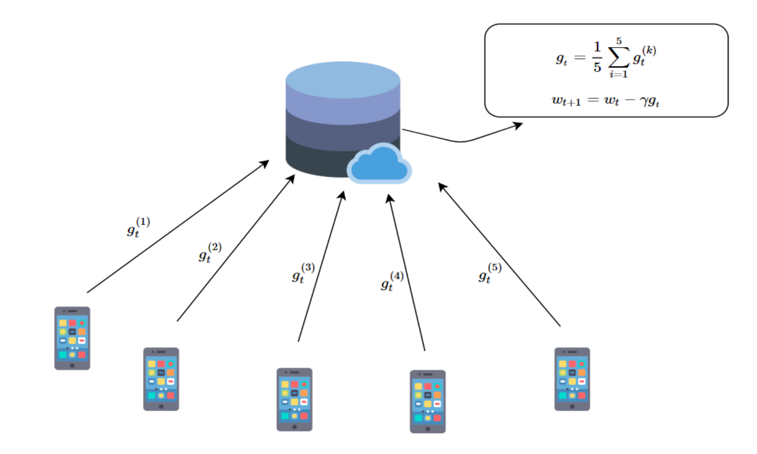

Figure: Sync-SGD (parameter upload)

Challenge?

What is the biggest challenge in Sync-SGD?

$\rightarrow$ Gradient communication between the parameter server and the worker causing bottleneck.

1. Gradient communication cost subsides gradient computation cost [1].

Way out: **Gradient compression**

[1] Yao *et al.*, 'Two-stream federated learning: Reduce the communication costs', *IEEE Visual Communications and Image Processing (VCIP)*, 2018.

LITERATURE SURVEY

1. Gradient sparsification, compression and quantization techniques have been introduced.
2. Communicate top-k gradients [2]. (sparsification)
3. Sign-SGD: 1-bit quantized gradients [3]. (quantization)
4. TernGrad: Quantize gradients to $\{-1, 0, 1\}$ [4]. (quantization)
4. Sketched SGD: Send the sketches of gradient [5]. (compression)

Is there a way to combine sparsification, compression and quantization?

$\rightarrow$ Quantized compressive sensing

[2] Stich *et al.*, 'Sparsified SGD with memory', *NIPS*, 2018.
[3] Bernstein *et al.*, 'signSGD: Compressed optimisation for non-convex problems', *ICML*, 2018.
[4] Wen *et al.*, 'TernGrad: Ternary Gradients to Reduce Communication in Distributed Deep Learning', *NIPS*, 2017.
[5] Ivkin *et al.*, 'Communication-efficient Distributed SGD with Sketching', *NIPS*, 2019.

Are gradients sparse?

1. Sparsity is induced by ReLU activation function.
   $f(x) = \max(0, x) \quad \rightarrow$ Forcing all $x < 0$ to 0
2. 44% of operations performed in most of the modern DNNs, for example AlexNet, GoogLeNet etc., are ineffective.

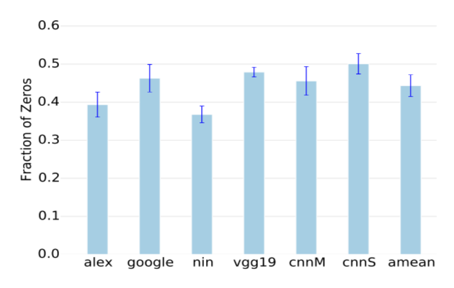

Figure: Average fraction of zero input neuron values in convolutional layer multiplication [6]

[6] Albericio *et al.*, 'Cnvlutin: Ineffectual-Neuron-Free Deep Neural Network Computing', *EEE ISCA*, 2016.

COMPRESSIVE SENSING

1. A sampling technique for signals which are sparse or compressible in some known basis [7].
2. Measurement matrix $\mathbf{\Phi}_{M\times N}$ is chosen to be a random matrix to obtain measurement vector $\mathbf{y}_{M\times 1}$ from signal $\mathbf{x}_{N\times 1}$ as,

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

3. Signal is recovered by solving LP optimization problem as follows:

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

[7] D. Donoho, 'Compressive Sensing', *IEEE Transactions on Information Theory*, 2006.

1. Quantization is modelled as as an additive measurement noise in quantized compressive sensing [8]: $\mathbf{y} = Q(\mathbf{\Phi x}) = \mathbf{\Phi x} + \mathbf{e}$

2. Measurement noise $\mathbf{n}$ is bounded by the quantization interval $\boldsymbol{\Delta}$ and the dimension of the compressed measurement (M): $||\mathbf{e}||_2 \leq \sqrt{\frac{M\Delta^2}{12}} = \epsilon$

Signal reconstructed by solving an optimization problem.

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

3. Reconstruction error $||\hat{\mathbf{x}} - \mathbf{x}||_2 = ||\mathbf{n}||_2 \leq \beta$

Issue?

$\rightarrow$ LP-based reconstruction is slow and computationally demanding.

[8] Boufounos *et al.*, '1-Bit compressive sensing', *42nd Annual Conference on Information Sciences and Systems.*, 2008.

PROPOSED APPROACH

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

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

3. Quantized compressed measurements $\mathbf{y}_t^{(k)}$ are sent to the parameter server.
4. At the parameter server the quantized compressed measurements are recovered to obtain $\tilde{\mathbf{g}}_t^{(k)}$.
5. Parameter server performs gradient aggregation: $\tilde{\mathbf{g}}_t = \frac{1}{K}\sum_{k=1}^{K} \tilde{\mathbf{g}}_t^{(k)}$.
6. 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

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

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

3. **(Variance bound assumption)** $\mathbb{E}[\mathbf{g}(\mathbf{w})] = \bar{\mathbf{g}}(\mathbf{w})$ and for some non-negative constant vector $\boldsymbol{\sigma} = [\sigma_1, \sigma_2, \ldots, \sigma_d]^T$,

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

4. Let $\bar{\mathbf{n}}_t = \mathbb{E}[\mathbf{n}_t]$ and there exists a non-negative μ such that,

$$\mu = \max_t \bar{\mathbf{g}}_t^T \bar{\mathbf{n}}_t$$

Theorem

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

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

1. SGD has same asymptotic convergence rate of $\mathbf{O}\left(\frac{\beta}{\sqrt{T}}\right)$ as of our approach.
2. TernGRAD provides probabilistic guarantee on convergence [9].
3. Error compensated DoubleSqueeze admits the same asymptotic convergence rate of $\mathbf{O}\left(\frac{\beta}{\sqrt{T}}\right)$ [10].

[9] Wen *et al.*, 'TernGrad: Ternary Gradients to Reduce Communication in Distributed Deep Learning', *NIPS*, 2017.
[10] Tang *et al.*, 'DoubleSqueeze: Parallel Stochastic Gradient Descent with Double-Pass Error-Compensated Compression', *Arxiv*, 2019.

COMPRESSIVE RECOVERY

Issues with LP-based CS recovery?

1. LP-based recovery algorithms are very slow $\rightarrow$ Slower convergence.
2. Large number of constraints $\rightarrow$ High computational complexity.

Way out?

$\rightarrow$ Iterative methods for CS recovery.

Advantages:

Identical to the LP-based CS recovery while running dramatically faster.

Restricted Isometry Property (RIP): Measurement matrix $\hat{\mathbf{\Phi}}$ holds RIP for all k-sparse signal $\mathbf{x}$ if,

$$(1-\delta_k)||\mathbf{x}||_2^2 \leq ||\hat{\mathbf{\Phi}}\mathbf{x}||_2^2 \leq (1+\delta_k)||\mathbf{x}||_2^2$$

Modified RIP: For $\mathbf{\Phi} = \frac{\hat{\mathbf{\Phi}}}{1+\delta_k}$ and $\beta_k = 1 - \frac{1-\delta_k}{1+\delta_k}$,

$$(1-\beta_k)||\mathbf{x}||_2^2 \leq ||\mathbf{\Phi}\mathbf{x}||_2^2 \leq ||\mathbf{x}||_2^2$$

Takeaway: $\mathbf{\Phi}$ holds RIP for sparsity k if $\beta_k < 1$.

Algorithm Definition Setting $\mathbf{x}_0 = 0$ for iteration $t = 0$,

$$\mathbf{x}_{t+1} = \mathcal{H}_k[\mathbf{x}_t + \mathbf{\Phi}^T(\mathbf{y} - \mathbf{\Phi}\mathbf{x}_t)]$$

where non-linear thresholding operator $\mathcal{H}_k(\mathbf{a})$ sets all but the largest k elements to 0.

1. Convergence is guaranteed when $\mathbf{\Phi}$ holds modified RIP.

1. Given noisy observation $\mathbf{y} = \mathbf{\Phi x} + \mathbf{e}$ ($\mathbf{x}$ being k-sparse) and $\mathbf{\Phi}$ maintaining modified RIP by $\beta_{3k} < \frac{1}{8}$, at t-th iteration we would obtain,

$$||\mathbf{x} - \mathbf{x}_t||_2 \leq 2^{-t}||\mathbf{x}_t||_2 + 4||\mathbf{e}||_2$$

2. Maximum number of iterations t^*,

$$t^* = \left\lceil \log_2 \frac{||\mathbf{x}||_2}{||\mathbf{e}||_2} \right\rceil$$

with accuracy $||\mathbf{x} - \mathbf{x}_{t^*}||_2 \leq 5||\mathbf{e}||_2$.

3. Complexity: $\mathbf{O}(t * \mathcal{L})$, where $\mathcal{L}$ denotes the complexity of applying $\mathbf{\Phi}$ and $\mathbf{\Phi}^T$.

Issue?

$\rightarrow$ Poor sparsity-undersampling tradeoff.

Recall: In IHT,

$$\begin{aligned} \mathbf{x}_{t+1} &= \mathcal{H}_k(\mathbf{\Phi}^T\mathbf{z}_t + \mathbf{x}_t) \\ \mathbf{z}_t &= \mathbf{y} - \mathbf{\Phi}\mathbf{x}_t \end{aligned}$$

AMP: Exploiting *belief propagation graphs*,

$$\begin{aligned} \mathbf{x}_{t+1} &= \mathcal{H}_k(\mathbf{\Phi}^T\mathbf{z}_t + \mathbf{x}_t) \\ \mathbf{z}_t &= \mathbf{y} - \mathbf{\Phi}\mathbf{x}_t + \frac{1}{\delta}\mathbf{z}_{t-1} < \mathcal{H}'_k(\mathbf{\Phi}^T\mathbf{z}_{t-1} + \mathbf{x}_{t-1}) > \end{aligned}$$

for $\mathbf{a} = [a(1), a(2), \ldots, a(N)]$, $< \mathbf{a} >= \sum_{i=1}^{N} \frac{a(i)}{N}$ and $\mathcal{H}'_k(\mathbf{s}) = \frac{\partial}{\partial \mathbf{s}}\mathcal{H}_k(\mathbf{s})$.

Extension to QCS

Recall: QCS can be modelled as $\mathbf{y} = Q(\mathbf{\Phi x}) = \mathbf{\Phi x} + \mathbf{e}$.

Idea?

$\rightarrow$ Consistency in measurement.

1. Minimize the loss function $\mathcal{C}$ defined as,

$$\mathcal{C}(\mathbf{y}, Q(\mathbf{\Phi \hat{x}}))$$

CONCLUSION

1. Combine compression and quantization $\rightarrow$ Quantized compressive sensing for gradient communication.
2. Convergence analysis for the proposed approach $\rightarrow$ Same with the asymptotic convergence rate of SGD: $\mathbf{O}(\frac{\beta}{\sqrt{T}})$.
3. In search of a new iterative QCS recovery algorithm $\rightarrow$ Combining with the idea of AMP.

BACKUP SLIDES

1. Form sketch of gradient $S(\mathbf{g}_t)$ of size $\mathbf{O}(\frac{1}{\epsilon} logN)$ to approximate gradient $\mathbf{g}_t$.
2. Recovery of gradient $\hat{\mathbf{g}}_t$ from $S(\mathbf{g}_t)$ fulfilling:

$$\mathbf{g}_i^2 - \epsilon||\mathbf{g}||_2^2 \leq \hat{\mathbf{g}}_i^2 \leq \mathbf{g}_i^2 + \epsilon||\mathbf{g}||_2^2 \tag{1}$$

3. ϵ is small error.
4. Sketched SGD approximating top-k gradients.

1. OMP has complexity of $\mathbf{O}(nmk)$.