BRAC UNIVERSITY

Inspiring Excellence

# **REPORT:** Near-optimal sparse allreduce for distributed deep learning

| | |
|---|---|
| **Name**: | Md Alkawser Rahman |
| **ID**: | 23366049 |
| **Course**: | CSE713 |
| **Semester**: | FALL2023 |

Based on: https://dl.acm.org/doi/abs/10.1145/3503221.3508399

**Paper Title:**

Near-optimal sparse allreduce for distributed deep learning

**Paper Link:**

**https://dl.acm.org/doi/abs/10.1145/3593013.3594095**

## 1 Summary

### 1.1 Motivation

The motivation behind this in addressing the critical impediment of communication overhead in distributed deep learning models. Recognizing the challenges of crafting efficient sparse allreduce algorithms and mitigating sparsification overhead, the report introduces the Ok-Topk scheme as a potential game-changer. The primary goal is to provide a comprehensive understanding of how Ok-Topk, with its revolutionary sparse allreduce algorithm and decentralized parallel Stochastic Gradient Descent (SGD) optimizer, offers a transformative solution. The paper seeks to motivate the broader deep learning community to reevaluate conventional approaches and embrace a future where communication efficiency and model training seamlessly coexist, combined with innovative contributions of the Ok-Topk scheme.

### 1.2 Contribution

The paper makes a significant contribution by introducing and extensively exploring the Ok-Topk scheme as a transformative solution to the challenge of communication overhead in distributed deep learning models. It emphasizes the scheme's effectiveness in reducing sparsification overhead while upholding convergence guarantees. The standout achievements, including superior scalability and a remarkable improvement for BERT on 256 GPUs, position Ok-Topk as a potential game-changer in large-scale model training, offering a paradigm shift in addressing communication efficiency challenges.

## 1.3 Methodology

### 1.3.1 Sparse Allreduce Algorithm:

The Ok-Topk scheme introduces a revolutionary sparse allreduce algorithm designed to be efficient and scalable, aiming to overcome the longstanding challenge in the distributed training landscape. It boasts a communication volume of less than 6k, representing an asymptotically optimal feat.

### 1.3.2 Gradient Sparsification:

Ok-Topk leverages gradient sparsification as a potent tool to alleviate the burdensome communication volume that comes with large-scale model training.The scheme efficiently selects the top-k gradient values guided by a meticulously estimated threshold, contributing to the reduction of sparsification overhead.

### 1.3.3 Decentralized Parallel Stochastic Gradient Descent (SGD) Optimizer:

Ok-Topk integrates its innovative sparse allreduce algorithm with the decentralized parallel Stochastic Gradient Descent (SGD) optimizer.This integration aims to fortify the solution with theoretical convergence guarantees, ensuring the effectiveness of the overall methodology in optimizing large-scale model training.

### 1.3.4 Threshold-Based Gradient Selection:

In the context of sparsification overhead reduction, Ok-Topk employs a threshold-based approach to efficiently select the top-k gradient values.This methodology ensures that the sparsification overhead is reduced without compromising the convergence guarantees fundamental to deep learning model training.

### 1.4 Conclusion

In a nutshell, the Ok-Topk scheme is positioned as more than just a solution; it represents a paradigm shift in addressing the formidable challenge of communication overhead. The narrative woven around innovation, efficiency, and practicality is supported by the integration of a pioneering sparse allreduce algorithm with decentralized parallel SGD. The paper concludes by urging the community to embrace a future where communication efficiency and model training seamlessly coexist, thanks to the Ok-Topk scheme presented by Li and Hoefler.

## 2. Limitation

### 2.1 Sensitivity to Hyperparameters:

One potential limitation could be the sensitivity of the Ok-Topk scheme to hyperparameters. The performance of the algorithm may be influenced by the choice of parameters such as the threshold for gradient selection or specific settings in the decentralized parallel Stochastic Gradient Descent (SGD) optimizer. Tuning these hyperparameters might be non-trivial and could impact the overall efficiency of the scheme.

### 2.2 Application Domain Specificity:

The effectiveness of the Ok-Topk scheme might be more pronounced in certain deep learning domains or specific architectures. There could be limitations in its generalizability across a wide range of applications, and its performance may vary based on the characteristics of the neural network models used. Understanding the scheme's domain-specific strengths and weaknesses is crucial for its practical applicability.

### 2.3 Computational Overhead during Gradient Selection:

The process of efficiently selecting the top-k gradient values to reduce sparsification overhead might introduce computational overhead. Depending on the size and complexity of the model, the overhead associated with this step could become a limiting factor in achieving real-time or near-real-time performance. This potential limitation needs to be considered, especially in scenarios where computational resources are constrained.

## 3 Synthesis

The introduction of Ok-Topk scheme as a potential game-changer, the report adeptly identifies and addresses the daunting hurdles of crafting an efficient sparse allreduce algorithm and mitigating the often underestimated sparsification overhead. The scheme's innovative integration of sparse allreduce algorithm with a communication coupled with the decentralized parallel Stochastic Gradient Descent (SGD) optimizer, sets the stage for a paradigm shift in large-scale model training. Ok-Topk's genius lies in its ability to efficiently select top-k gradient values, reducing sparsification overhead while upholding convergence guarantees. Thus, showcasing its superiority over existing methodologies and culminating in a call to embrace a future where communication efficiency and model training seamlessly coexist.