COMET: A Novel Memory-Efficient Deep Learning Training Framework by Using Error-Bounded Lossy Compression

Sian Jin1, Chengming Zhang1, Xintong Jiang2, Yunhe Feng3, Hui Guan4, Guanpeng Li5, Shuaiwen Leon Song6, Dingwen Tao1
1Indiana University, 2McGill University, 3University of Washington, 4University of Massachusetts, 5University of Iowa, 6University of Sydney

Abstract

Training wide and deep neural networks require large amounts of storage resources such as memory because the intermediate activation data must be saved in the memory during forward propagation and then restored for backward propagation. However, state-of-the-art accelerators such as GPUs are only equipped with very limited memory capacities due to hardware design constraints, which significantly limits the maximum batch size and hence performance speedup when training large-scale DNNs. In this paper, we propose a novel memory-efficient CNN training framework (called COMET) that leverages error-bounded lossy compression to significantly reduce the memory requirement for training in order to allow training larger models or to accelerate training. Our framework purposely adopts error-bounded lossy compression with a strict error-controlling mechanism.

Contribution

This work has mainly five contributions:
1. A novel memory-efficient CNN training framework via dynamically compressing the intermediate activation data through error-bounded lossy compression
2. A thorough analysis of the impact of compression error propagation during DNN training from both theoretical and empirical perspectives
3. An adaptive scheme to adaptively configure the error-bounded lossy compression based on a series of current training statistics
4. Improved SZ error-bounded lossy compression to handle compressing continuous zeros
5. Reduce the memory consumption by up to $13.5\times$ and $1.8\times$ compared to the original training framework and the state-of-the-art method, respectively. Improve the end-to-end training performance by up to $2\times$

Methodology

- Memory consumption and top-1 accuracy of different state-of-the-art neural networks: large memory consumption

Evaluation

- Comparison of accuracy and activation size between baseline training and our proposed framework
- Accuracy curve comparison between the baseline and our proposed framework: high compression ratio with almost no accuracy degradation

Parameter Collection:
- collect parameters for analysis and updating compression configurations

Gradient Assessment:
- estimate acceptable variance in the gradient

Activation Assessment:
- estimate acceptable error introduced for compressing activation data

Adaptive Compression:
- deploy lossy compression according to current training status data

- An accurate theoretical prediction to the gradient error distribution

Next Layer

Forward

- Conv

- Activation

- Data

- Sparsity

- Max

- Loss

Average

Backward

- Loss

- Data

- Flow

- Value

Flow

- Value

Flow

Per

1000

Iterations

- Accurate theoretical prediction to the gradient error distribution

Overview of our proposed memory-efficient DNN training framework, COMET

1. Parameter Collection: collect parameters for analysis and updating compression configurations
2. Gradient Assessment: estimate acceptable variance in the gradient
3. Activation Assessment: estimate acceptable error introduced for compressing activation data
4. Adaptive Compression: deploy lossy compression