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

INDIANA UNIVERSITY BLOOMINGTON

Sian Jin[1], Chengming Zhang[1], Xintong Jiang[2], Yunhe Feng[3], Hui Guan[4], Guanpeng Li[5], Shuaiwen Leon Song[6], Dingwen Tao[1]

[1]Indiana University, [2]McGill University, [3]University of Washington, [4]University of Massachusetts, [5]University of Iowa, [6]University of Sydney

## Abstract

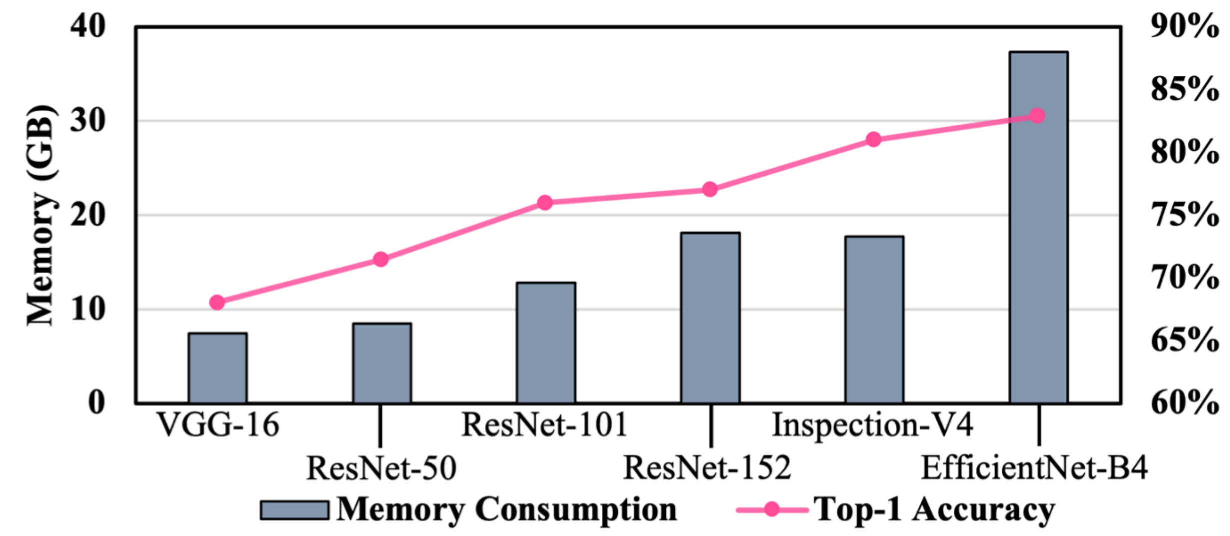
Training wide and deep neural networks require large amounts of storage resources such as memory because the intermediate activation data must be saved in the memory during forward propagation and then restored for backward propagation. However, state-of-the-art accelerators such as GPUs are only equipped 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

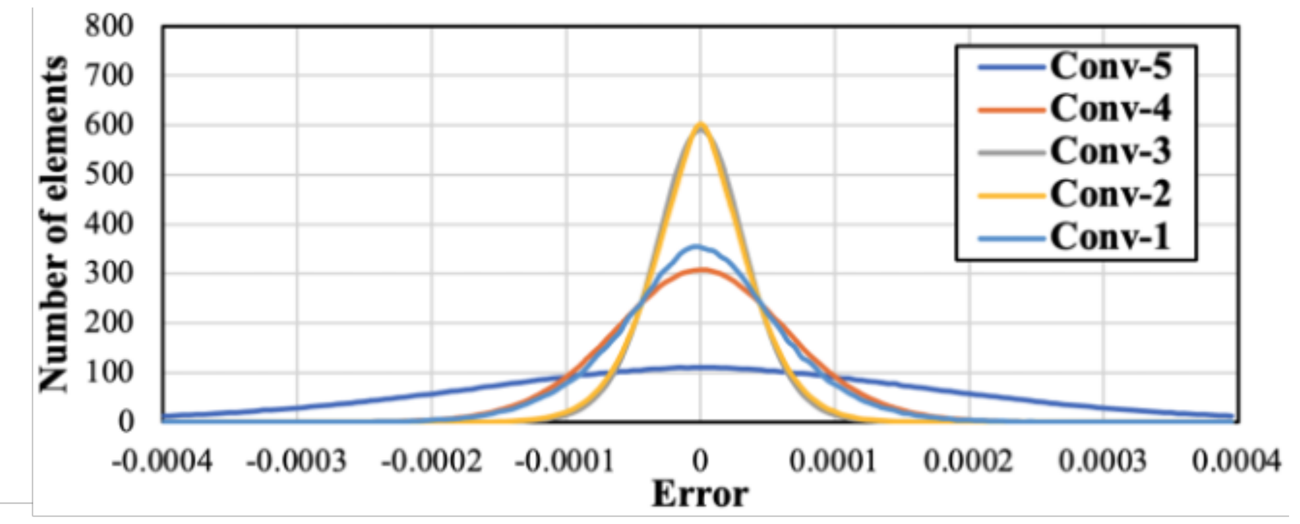
This work has mainly five contributions:

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 status data
4. Improved SZ error-bounded lossy compression to handle compressing continuous zeros
5. Reduce the memory consumption by up to 13.5× and 1.8× 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×

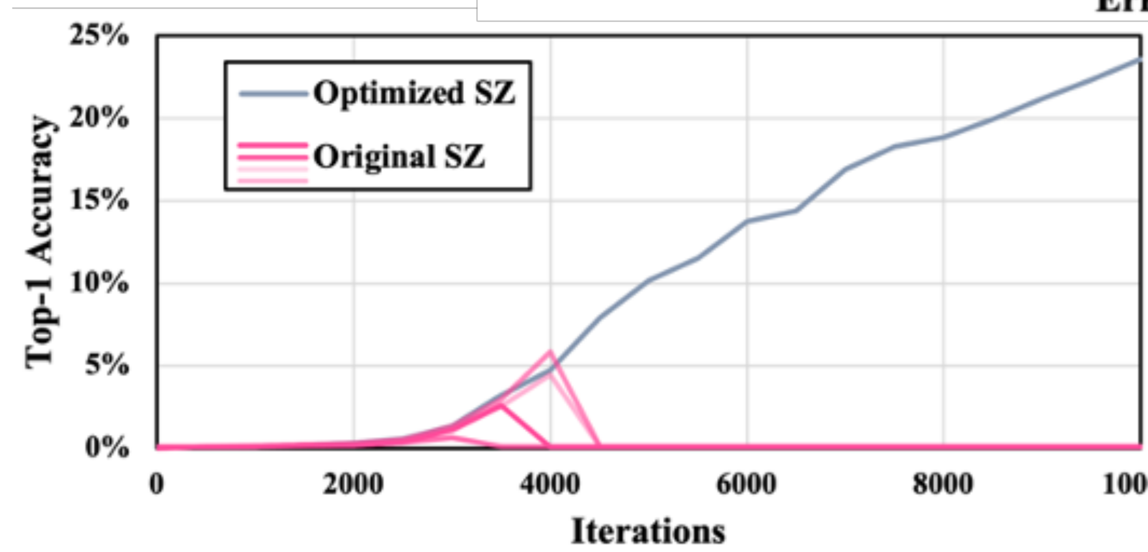
## Methodology




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

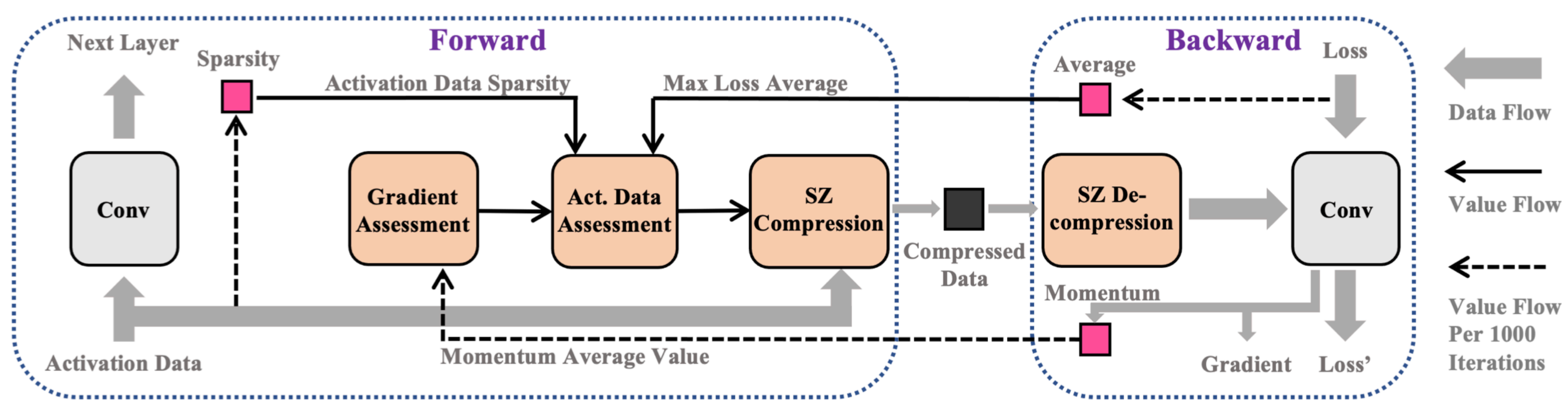



→ An example of gradient error distribution after compression: normal distribution, consistent with our prediction




← Training failed without modification to the compressor: need modified compressor

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



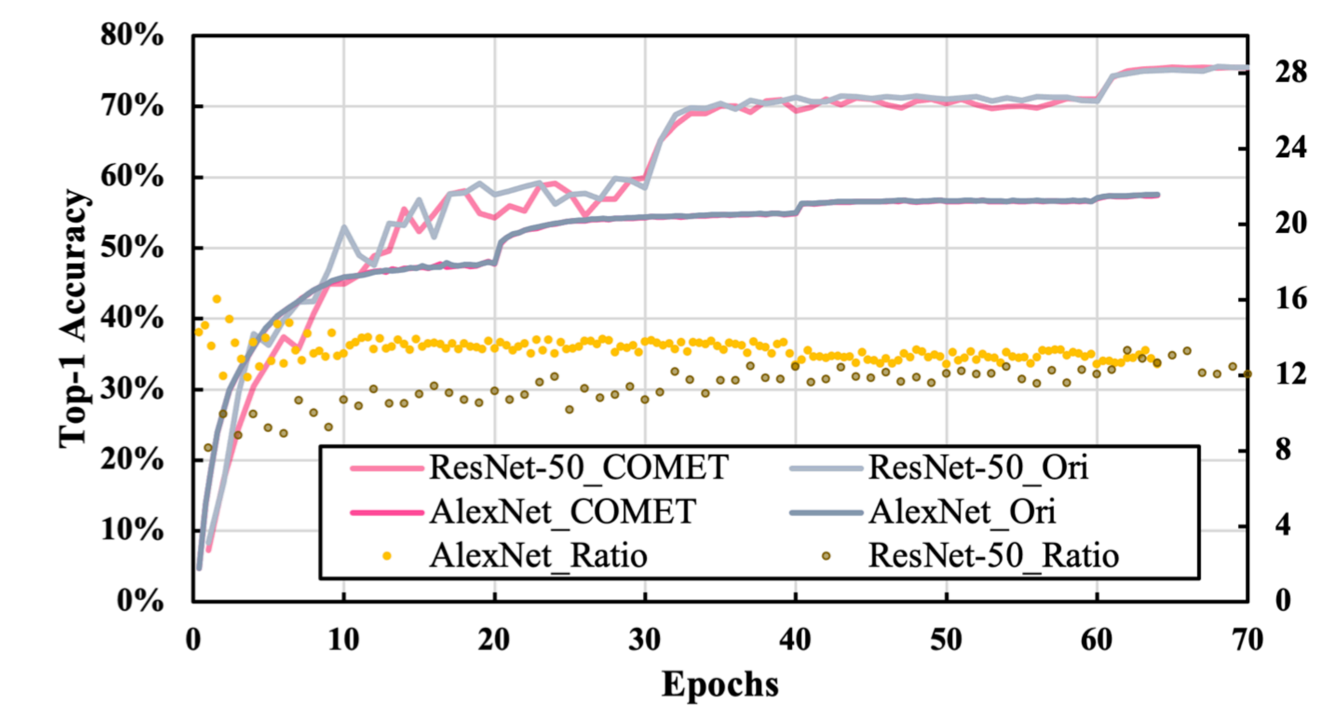

↑ Overview of our proposed memory-efficient DNN training framework, COMET

## Evaluation

| Neural Nets | | Top-1 Accuracy | Peak Mem. | Max Batch | Conv. Act. Size | COMET | JPEG-ACT |
|---|---|---|---|---|---|---|---|
| AlexNet | b. | 57.41% | 2.17 GB | 512 | 407 MB | 13.5× | - |
| | c. | 57.42% | 0.85 GB | 2048 | **30 MB** | | |
| VGG-16 | b. | 68.05% | 17.29 GB | 64 | 6.91 GB | 11.1 × | - |
| | c. | 68.02% | 5.04 GB | 256 | **0.62 GB** | | |
| ResNet-18 | b. | 67.57% | 5.16 GB | 256 | 1.71 GB | 10.7 × | 7.3 × |
| | c. | 67.43% | 1.37 GB | 1024 | **0.16 GB** | | |
| ResNet-50 | b. | 75.55% | 15.57 GB | 128 | 5.14 GB | 11.0 × | 6.0 × |
| | c. | 75.51% | 4.40 GB | 512 | **0.46 GB** | | |

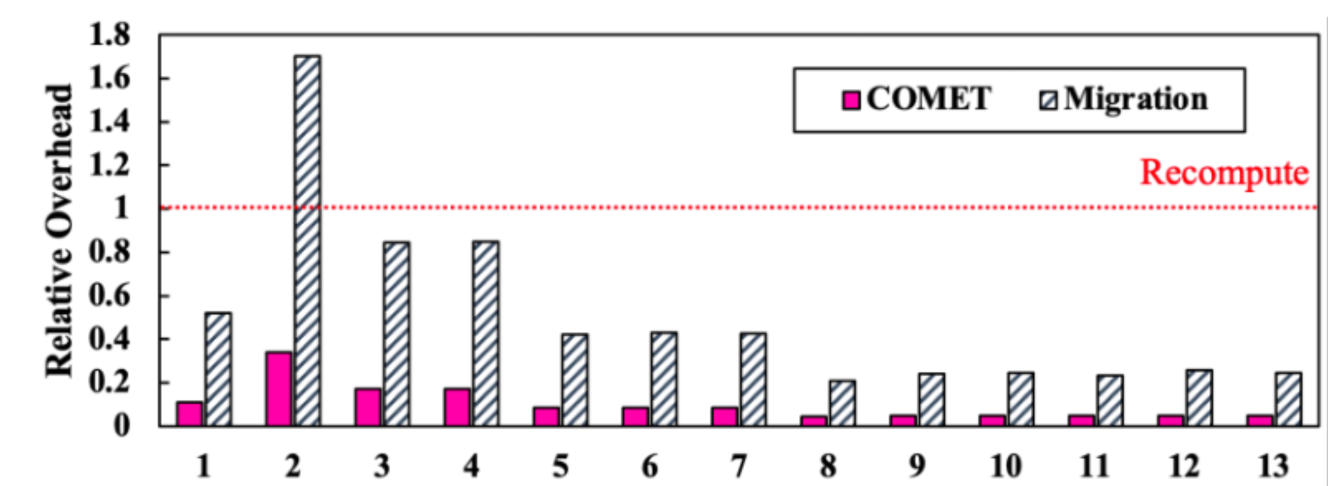
b.= baseline, c.= compressed

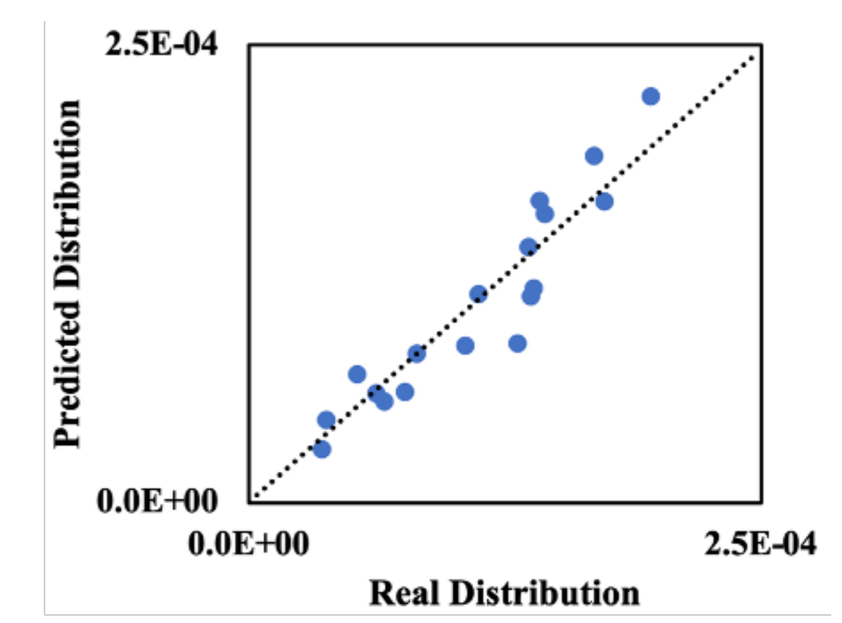
← Comparison of accuracy and activation size between baseline training and our proposed framework




← Training accuracy curve comparison between the baseline and our proposed framework: high compression ratio with almost no accuracy degradation

→ Overhead comparison between migration, recomputation: our solution provides the lowest overhead




↑ Accurate theoretical prediction to the gradient error distribution