POSTER: A Novel Memory-Efficient Deep Learning Training Framework via Error-Bounded Lossy Compression

Sian Jin
Washington State University
Pullman, WA, USA
sian.jin@wsu.edu

Guanpeng Li
University of Iowa
Iowa City, IA, USA
guanpeng-li@uiowa.edu

Shuaiwen Leon Song
University of Sydney
Sydney, NSW, Australia
shuaiwen.song@sydney.edu.au

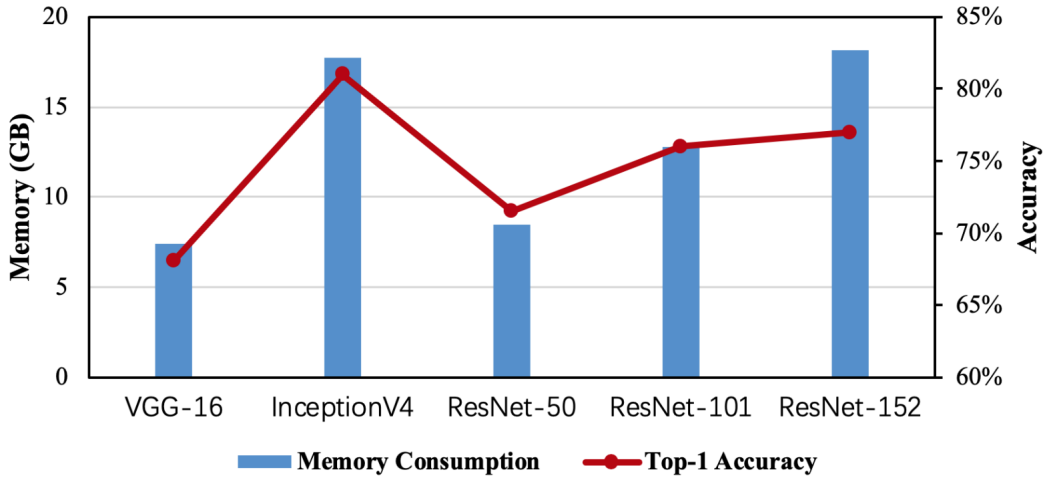
Dingwen Tao*
Washington State University
Pullman, WA, USA
dingwen.tao@wsu.edu

2021

Why Compress DNNs' Activation Data

WASHINGTON STATE UNIVERSITY THE UNIVERSITY OF IOWA THE UNIVERSITY OF SYDNEY

- **Why Compress?**
  - **High memory consumption**
  - Large batch size needed
  - Highly limited GPU memory space

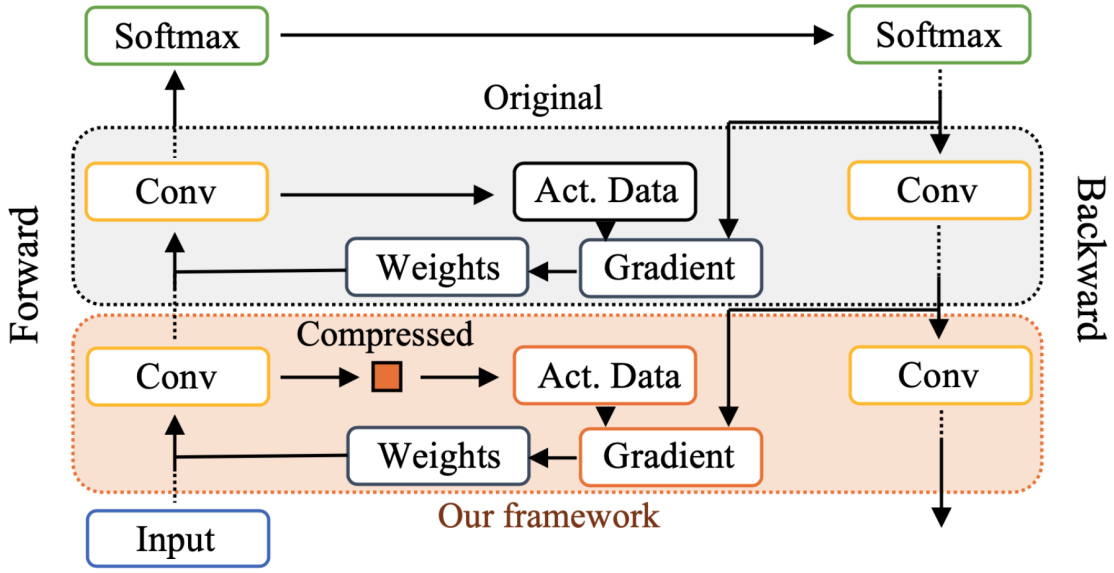


Memory consumption and top-1 accuracy of different state-of-the-art neural networks.

Deep Neural Networks

- **Why Compress?**
  - **High memory consumption**
  - Large batch size needed
  - Highly limited GPU memory space

- **Activation Data Storage For Training**
  - Must stored until used in back propagation
  - Long waiting period between generating and using the data



Data flow in a sample iteration of training CNNs

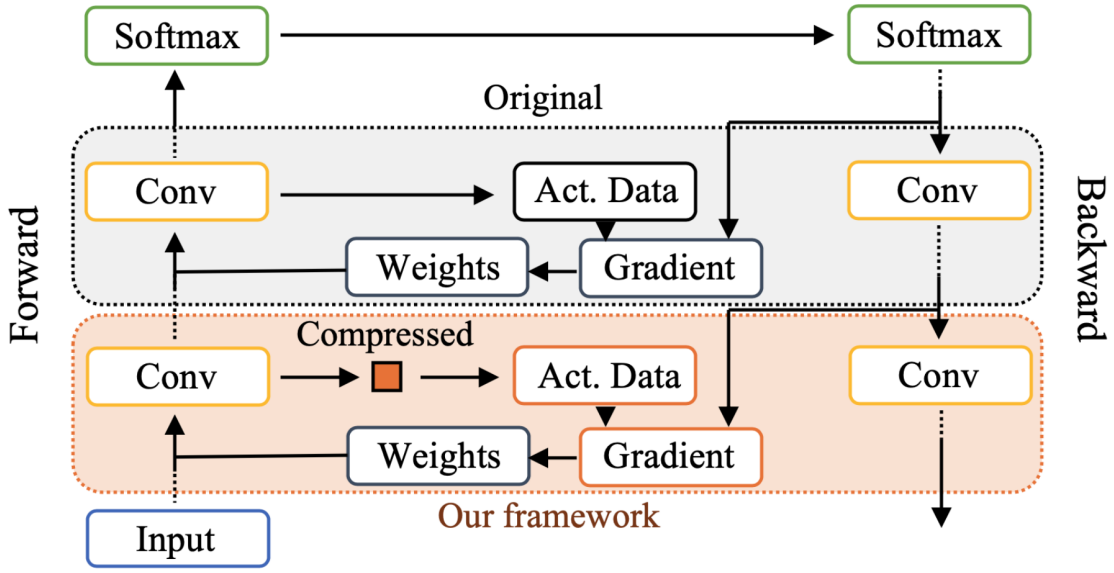
Deep Neural Networks

➢ **Activation Data Storage For Training**

- Must stored until used in back propagation
- Long waiting period between generating and using the data

➢ **Convolutional Layer?**

- Not able to recompute without high overhead
- Large proportion of activation data consumption



Data flow in a sample iteration of training CNNs

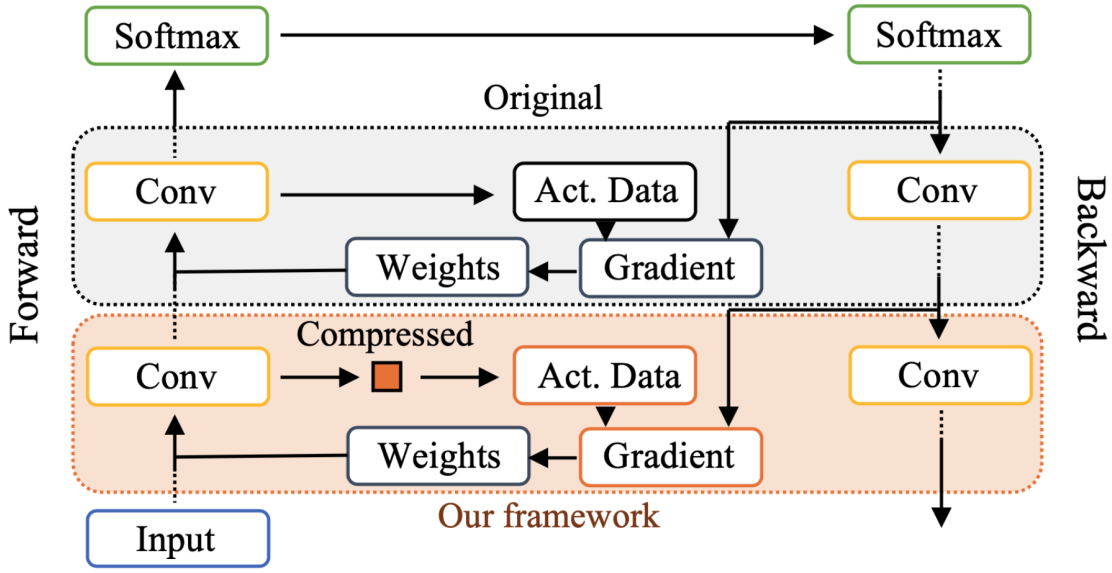
Deep Neural Networks

➢ Convolutional Layer?

- Not able to recompute without high overhead
- Large proportion of activation data consumption

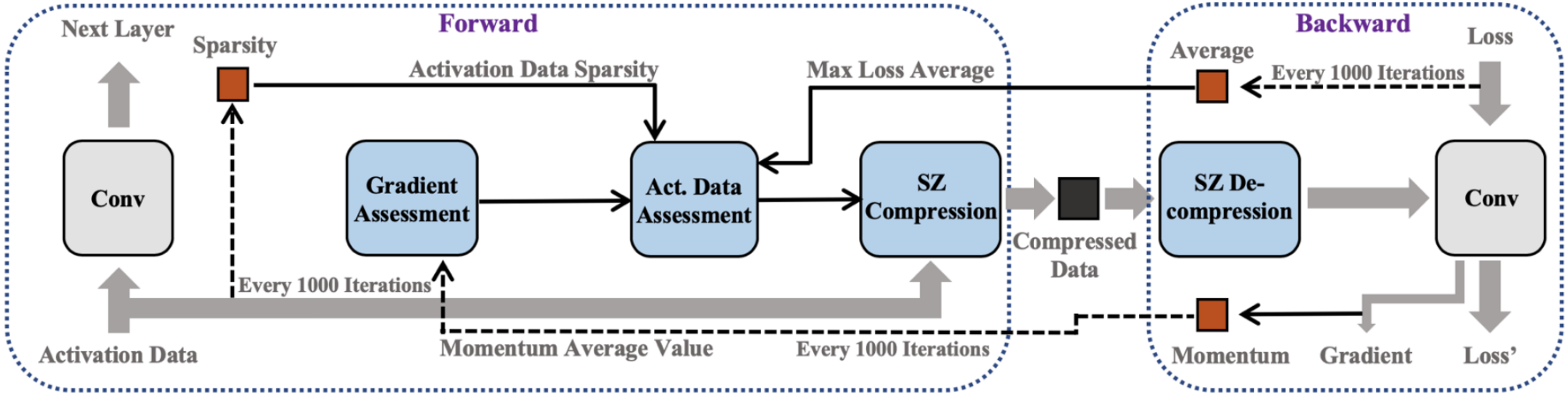
➢ **Why SZ Lossy Compression?**

- **High compression ratio** compared to lossless compression
- Specifically designed for floating-point data
- GPU enabled, with **high (de)compression throughput**



Data flow in a sample iteration of training CNNs

Overall Design



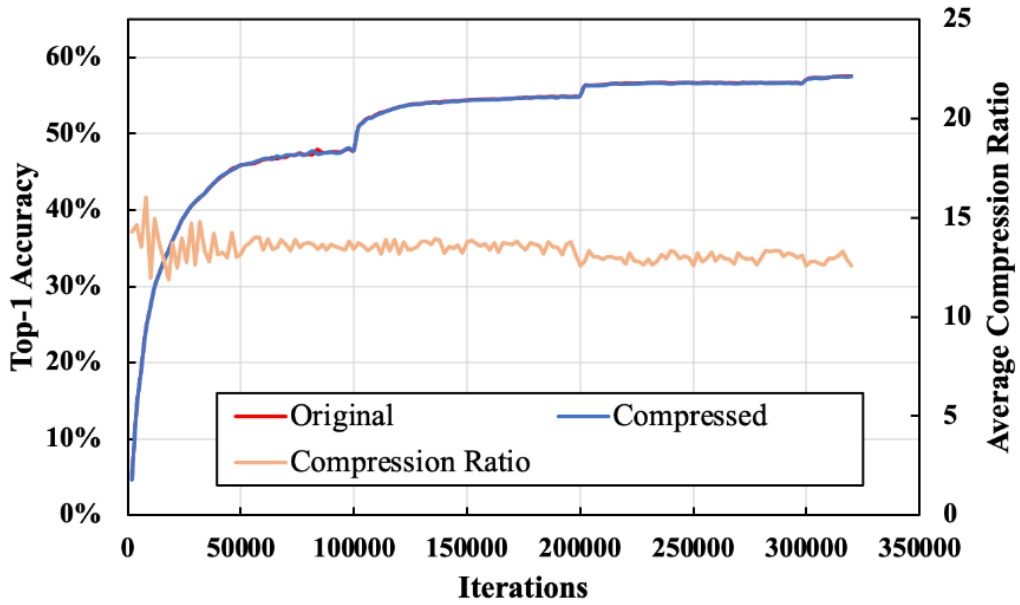
Overview of our proposed memory-efficient DNN training framework

- **Parameter Collection**: collect parameters for analysis and updating compression configurations
- **Gradient Assessment**: estimate acceptable σ in the gradient
- **Activation Assessment**: estimate acceptable error bound for compressing activation data
- **Adaptive Compression**: deploy lossy compression

Evaluation

➢ Memory Reduction Evaluation

- **High compression ratio, up to 13.5x**
- **Little/no testing accuracy loss**



Training accuracy curve comparison between the baseline and our proposed framework.

Neural Network	Top-1 Accuracy	Convolutional Act. Size	Compress Ratio
baseline	57.41%	407 MB	
AlexNet compressed	57.42%	**30 MB**	13.5×
baseline	68.05%	9.30 GB	
VGG-16 compressed	68.02%	**0.83 GB**	11.1 ×
baseline	67.57%	3.42 GB	
ResNet-18 compressed	67.43%	**0.32 GB**	10.7 ×
baseline	71.49%	10.28 GB	
ResNet-50 compressed	71.18%	**0.93 GB**	11.0 ×

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

Evaluation

➢ Memory Reduction Evaluation

- **High compression ratio, up to 13.5x**
- **Little/no testing accuracy loss**

➢ Performance Evaluation and Analysis

- Low introduced overhead, as low as **7%** of total computing time
- High performance improvement, up to **1.27x** raw performance

Neural Network	Top-1 Accuracy	Convolutional Act. Size	Compress Ratio
baseline	57.41%	407 MB	
AlexNet compressed	57.42%	**30 MB**	13.5×
baseline	68.05%	9.30 GB	
VGG-16 compressed	68.02%	**0.83 GB**	11.1 ×
baseline	67.57%	3.42 GB	
ResNet-18 compressed	67.43%	**0.32 GB**	10.7 ×
baseline	71.49%	10.28 GB	
ResNet-50 compressed	71.18%	**0.93 GB**	11.0 ×

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

Conclusion and Future Work

WASHINGTON STATE UNIVERSITY THE UNIVERSITY OF IOWA THE UNIVERSITY OF SYDNEY

- **Our Proposed Framework**
  - A novel memory-efficient DNN training framework via dynamically compressing the intermediate activation data through error-bounded lossy compression
  - A thorough analysis on the impact of compression error propagation during CNN training from both theoretical and empirical perspectives
  - An adaptive scheme to dynamically configure the error-bounded lossy compression based on a series of current training status data
  - Evaluate our proposed framework on popular models
    - Can provide up to **13.5x** compression ratio, up to **1.8x** improvement compared to previous state-of-the-art method
- **Future Work**
  - Implement orthogonal methods including data migration and layer re-computation
  - Optimize (de)compression performance on activation data

Thank you!

Any questions are welcome!

Contact Dingwen Tao: dingwen.tao@wsu.edu
Sian Jin: sian.jin@wsu.edu