# COAT: Compressing Optimizer states and Activation for Memory-Efficient FP8 Training

Haocheng Xi[1], Han Cai[2], Ligeng Zhu[2], Yao Lu[2], Kurt Keutzer[1], Jianfei Chen[4], Song Han[23]

[1] University of California, Berkeley [2] NVIDIA [3] MIT [4] Tsinghua University

## Overview

COAT is a Memory Efficient FP8 Training technique:

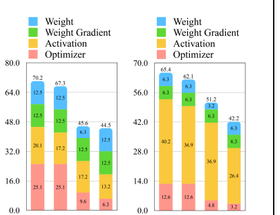
- **FP8 Optimizer states** → 4× memory reduction
- **FP8 Activations** → 2× memory reduction
- End-to-End **1.43× Speedup**
- End-to-End **1.54x memory reduction**

Github Arxiv

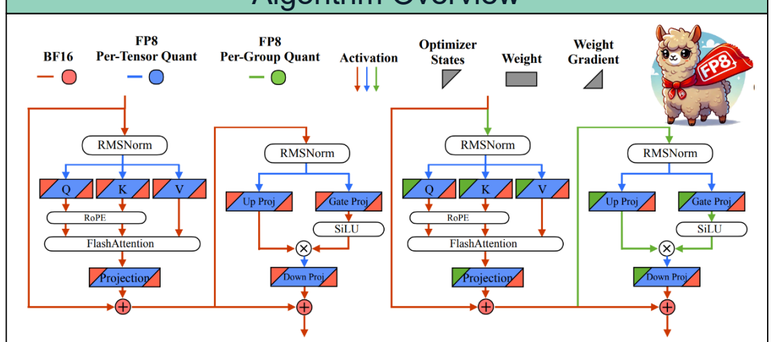
## Motivation

Training LLMs is very memory intensive.
Train a 7B LLM requires at least 120GB memory!

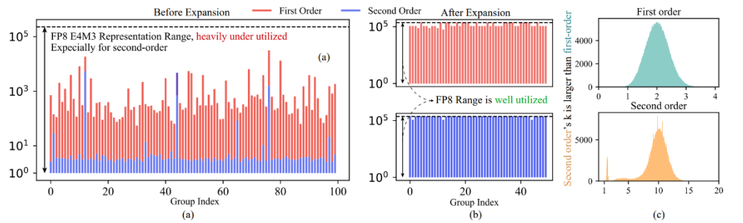
- Adam **Optimizer States** → 8 bytes / param
  - First Order Momemtum (4 bytes)
  - Second Order Momentum (4 bytes)
- **Activation** → Proportional to batch size
  - Required by backward pass
  - usually > 30GB / GPU




## Algorithm Overview




## FP8 Optimizer




**Problem**: Optimizer states **waste** the representation ability.
Its **dynamic range** is much smaller than FP8's dynamic range

$$\mathcal{R}_{f(X)} = \frac{\max(|f(X)|)}{\min(|f(X)|)} = \frac{\max(|\text{sign}(X)X^k|)}{\min(|\text{sign}(X)X^k|)} = \left(\frac{\max(|X|)}{\min(|X|)}\right)^k = (\mathcal{R}_X)^k.$$

$$\mathcal{R}_X = \frac{\max(|x_1|, |x_2|, \ldots, |x_G|)}{\min(|x_1|, |x_2|, \ldots, |x_G|)}$$

Dynamic Range

**Solution**: Dynamic range expansion function makes the distribution better

$$f(x) = \text{sign}(x)|x|^k$$

| MSE of $\frac{m}{\sqrt{v}}$ | Second Order | | | |
|---|---|---|---|---|
| **First Order** | **E4M3** | **E4M3+Expand** | **E5M2** | **E5M2+Expand** |
| **E4M3** | **20.10** | 18.08 | **25.65** | 18.16 |
| **E4M3+Expand** | 15.13 | **12.31** | 21.96 | **12.43** |
| **E5M2** | **37.02** | 35.96 | **40.30** | 36.00 |
| **E5M2+Expand** | 17.79 | 15.48 | 23.84 | 15.57 |


## FP8 Activation

Observation:
**Non-linear layers** takes >50% of the Activation Memory.

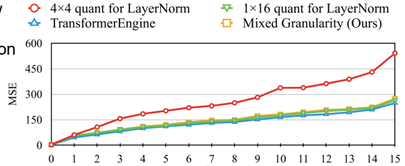
| | | Non-Linear | | Attention | | | | Reduction Ratio | |
|---|---|---|---|---|---|---|---|---|---|
| | | RMSNorm | Act Func | RoPE | FlashAttn | Linear | Total | Ideal | Achieved |
| Llama-style | BF16 | 4U | 8U | 2U | 3U | 5.66U | 22.66U | 1.00× | 1.00× |
| | TE | 2U | 8U | 2U | 3U | 3.33U | 18.33U | 1.23× | 1.20× |
| | COAT | 1U | 4U | 2U | 3U | 3.33U | 13.33U | 1.69× | 1.65× |

**Method**: Mixed granularity FP8 precision flow

4×4 quant for LayerNorm
TransformerEngine
1×16 quant for LayerNorm
Mixed Granularity (Ours)

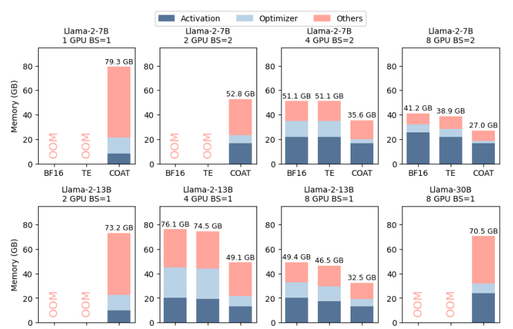
Nonlinear layers: fine-grained quantization
Linear layer: Per-tensor Quantization
More friendly with FP8 TensorCore

Store FP8 tensors for backward
50% memory reduction!



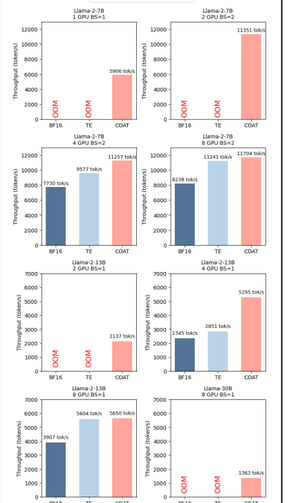

## Memory Reduction

**1.54×** End-to-End Memory Reduction!
**Double** the maximum batch size & **Half** the required device!



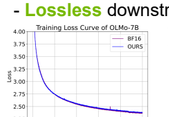

## Training Speedup

**1.45×** End-to-End Speedup in average!
**2.25×** End-to-End Speedup maximum!




## Lossless Performance

We train a 7B LLM with 300B token. We achieve:
- Training curve **matches** exactly
- **Lossless** downstream task performance




| | Train Loss | WikiText | C4 | Pile | Avg ppl |
|---|---|---|---|---|---|
| **BF16** | 2.366 | 12.053 | 12.874 | 8.596 | 11.174 |
| **COAT** | 2.379 | 12.166 | 12.988 | 8.684 | 11.279 |

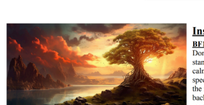
| | COPA | ARC(Easy) | SciQ | HellaSwag | Avg Acc |
|---|---|---|---|---|---|
| **BF16** | 83.0% | 65.7% | 87.5% | 56.9% | 73.2 % |
| **COAT** | 81.0% | 61.9% | 87.2% | 60.6% | 72.7 % |

We fine-tune a Vision Language Model.
- Performance **on-par** with BF16 training
- Application quality is maintained


| Stage 3 | VideoMME | POPE | VizWiz | GQA* | VQAv2* |
|---|---|---|---|---|---|
| BF16 | 42.96 | 86.90 | 61.42 | 64.55 | 81.47 |
| TE | 43.19 | 87.64 | 57.61 | 64.53 | 81.34 |
| COAT | 44.56 | 87.43 | 61.36 | 64.44 | 81.20 |

SEED

| Stage 3 | TextVQA | Image | Video | MMMU Val | Average |
|---|---|---|---|---|---|
| BF16 | 65.60 | 73.40 | 45.65 | 38.56 | 62.80 |
| TE | 64.70 | 73.51 | 43.12 | 35.89 | 61.88 |
| COAT | 64.65 | 73.36 | 43.76 | 37.22 | 62.51 |



**Instruction: Briefly describe this figure.**

**BF16**: The image depicts a fantastical scene set in a mountainous landscape. Dominating the center of the image is a large tree, its twisted and gnarled form standing on a rocky outcropping. The tree is surrounded by a serene lake, its calm waters reflecting the vibrant colors of the sky above. The sky is a spectacle in itself, painted with hues of orange and red as the sun sets behind the mountains, casting a warm glow over the entire scene. The mountains in the background add a sense of depth and grandeur to the image, their peaks reaching up towards the sky. The overall composition of the image suggests a tranquil yet powerful natural setting, possibly inspired by mythology or fantasy.

**COAT**: The image presents a captivating scene of a solitary tree standing on a rocky outcropping. The tree, with its twisted and gnarled form, is a vibrant green, contrasting with the surrounding landscape. The rocky outcropping on which it stands is rugged and uneven, adding to the overall wildness of the scene. The background is a breathtaking view of a mountainous landscape under a sky ablaze with hues of orange and red, possibly indicating a sunrise or sunset. The sun is partially obscured by the clouds, casting a warm glow that illuminates the scene. The image is rendered in a digital art style, which adds a sense of realism and depth to the scene. The artist has skillfully used color and lighting to create a sense of tranquility and solitude, with the tree standing as a solitary figure against the vastness of the landscape. The overall composition of the image is balanced and harmonious, with each element complementing the others to create a cohesive whole.