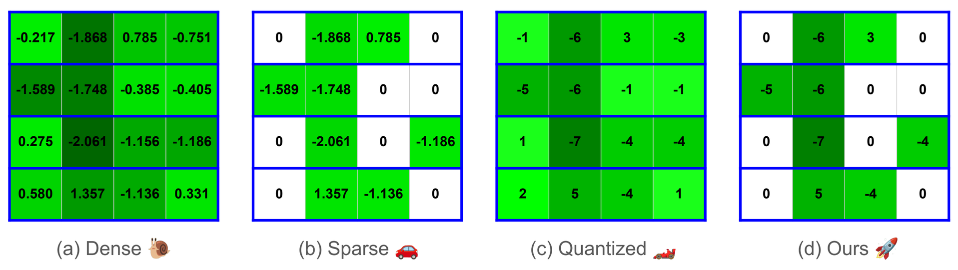
# SLOPE: Extreme Model Compression with Structured Sparsity at Low Precision

Dan Liu[1,2] Nikita Dvornik[2] Xue Liu[1,2,3]

[1]McGill University [2]Palona AI [3]MBZUAI

## Observation

While both quantization and structured sparsity are effective for model compression, directly combining them often causes severe accuracy degradation due to compounded distortions in both magnitude and direction of the weights.




- Two dominant compression strategies:
  - **Quantization** reduces numerical precision (e.g., 32-bit → 4-bit).
  - **Structured $N{:}M$ sparsity** prunes redundant weights (e.g., 2:4).
- Each works well individually, but **naïve combination** causes compounding distortions (magnitude + direction), leading to large accuracy drops.
- We observe severe **angular deviation** and **SQNR** degradation when applying sparsity and quantization together.

| Setting | Acc. (%) | Cos. ↑ | SQNR (dB) ↑ |
|---|---|---|---|
| FP32 | 69.76 | – | – |
| 2:4 Sparse | 69.96 | 0.975 | 14.7 |
| 4-bit Quant | 71.10 | 0.948 | 10.5 |
| 2:4 + 4-bit | **68.36** | **0.919** | **8.4** |

Empirically, ResNet-18 accuracy drops from 69.96% (2:4 sparse) or 71.10% (4-bit) to only 68.36% when both are applied. Angular deviation (cosine 0.919) and SQNR (8.4 dB) confirm compounded distortion, motivating an angular-aware regularizer.

## Motivation and Goal

The goal of SLOPE is to achieve **extreme model compression** by jointly applying structured $N{:}M$ sparsity and low-bit quantization while preserving accuracy.

- **Observation:** magnitude-based losses (L2) fail to recover accuracy under sparse quantization.
- **Insight:** preserving *directional alignment* between full-precision and sparse-quantized weights stabilizes training.
- **Goal:** 20× compression with >99 % accuracy retention.

## Key Contribution — SLOPE Regularizer

We propose a **cosine-based regularization** that enforces angular consistency:

$$\min_{W} \mathcal{L}(\hat{W}) + \lambda \mathcal{L}_{\text{reg}}, \qquad \mathcal{L}_{\text{reg}} = \frac{1}{n}\sum_{i}(1 - \cos(w_i, \hat{w}_i)).$$

- Minimizes angular deviation instead of element-wise difference.
- Theoretical bound: $\|w - \hat{w}\|_2^2 \le 2\|w\|_2^2(1 - \cos\theta)$.
- Implemented with **Straight-Through Estimator (STE)** for differentiable training.
- Compatible with **NVIDIA 2:4 sparsity kernels**.

## Method and Training Pipeline

SLOPE integrates quantization $q(\cdot)$ and sparsification $S(\cdot)$:

$$\hat{W} = q(S(W; N, M); s, b),$$

where $s$ is a learnable scale, $b$ the bit-width, and $N{:}M$ the structured sparsity pattern.

1. Apply structured sparsity ($N{:}M$) to weights.
2. Quantize remaining non-zeros to $b$-bit precision (LSQ formalism).
3. Optimize full-precision proxy $W$ using STE: $W_{t+1} = W_t - \eta\, \partial\mathcal{L}(\hat{W}_t)/\partial\hat{W}_t$.
4. Regularizer aligns directions of $W$ and $\hat{W}$.

## Results — Image Classification (ImageNet)

| Model | Setting | Baseline (%) | SLOPE (%) | Gain |
|---|---|---|---|---|
| ResNet-18 | 2:4 + 4-bit | 68.36 | **71.11** | +2.75 |
| ResNet-18 | 2:4 + 2-bit | 64.47 | **67.59** | +3.12 |
| ResNet-50 | 2:4 + 4-bit | 74.70 | **75.93** | +1.23 |
| DeiT-Small | A16/W4 | 79.28 | **80.41** | +1.13 |

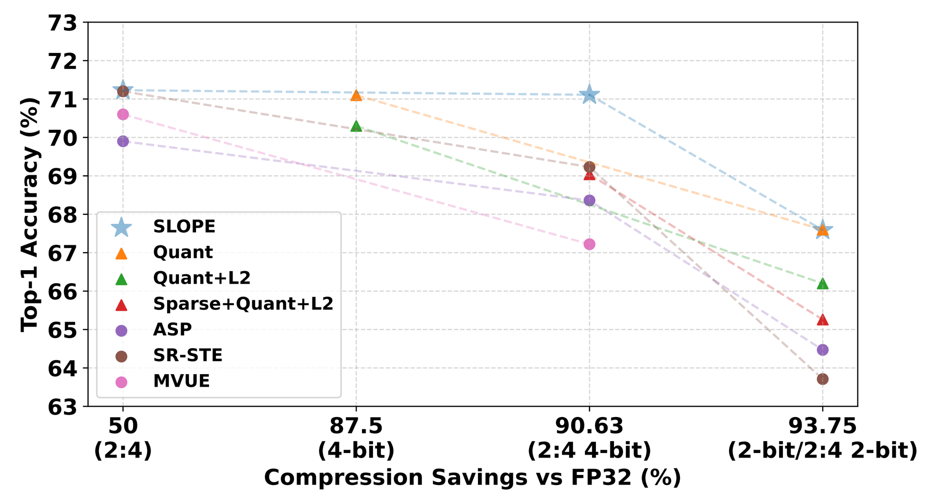



Figure 1. Accuracy vs. compression ratio on ResNet-18 models.

- Up to **20×** model-size reduction.
- SLOPE maintains ∼99% of FP32 accuracy under 2–4 bit quantization.

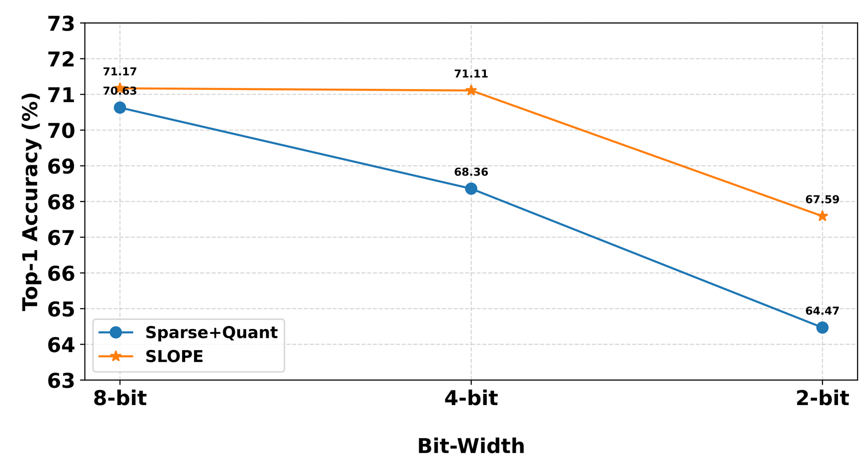



Figure 2. Accuracy of 2:4 Sparse+Quant and SLOPE under varying bit-widths on ResNet-18.

## Results — Object Detection & Segmentation (COCO)

| Setting | Box mAP | Mask mAP |
|---|---|---|
| Dense W4 | 37.80 | 34.92 |
| **SLOPE W4** | **40.83** | **37.13** |
| Dense A4/W4 | 29.64 | 27.59 |
| **SLOPE A4/W4** | **38.79** | **35.28** |

SLOPE improves both box and mask mAP by 2–3 points over dense baselines.

## Analysis and Ablation

### Effect of Sparsity Level

Accuracy remains stable even for 2:8 and 2:16 sparsity — SLOPE mitigates degradation by preserving weight directions.

### Comparison of Regularizers

| Setting | Baseline | +L2 | SLOPE (cos) |
|---|---|---|---|
| 4-bit 2:4 | 68.36 | 69.04 | **71.11** |
| 2-bit 2:4 | 64.47 | 65.26 | **67.59** |

Table 1. Directional regularization vs. L2 penalty.

Directional regularizer consistently outperforms L2, confirming angular alignment is key to accuracy recovery.

## Conclusion

- Unified framework for combining structured sparsity and quantization.
- Angular regularizer stabilizes optimization under extreme compression.
- Generalizes across CNNs, ViTs, and detection models.
- Enables 2–4 bit quantization with up to 93.75 % compression and near-lossless accuracy.


daniel.liu@mail.mcgill.ca