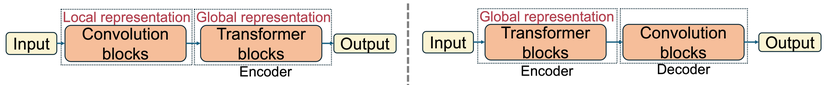
EfficientQuant: An Efficient Post-Training Quantization for CNN-Transformer Hybrid Models on Edge Devices

Shaibal Saha, Lanyu Xu
Oakland University
Contact: {shaibalsaha, lxu}@oakland.edu

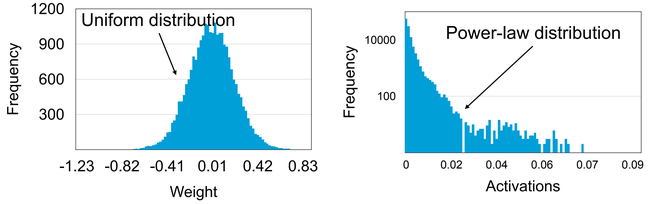
Motivation

➢ Different Representation of Hybrid Models



Automatic identification of convolutional and transformer blocks for quantization?

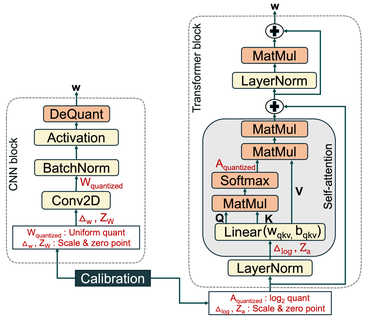
➢ Weight and Activation Distribution for Convolutional Block and Transformer block



Can one quantization method handle the diverse components of hybrid models?

EfficientQuant: Structure-Aware Quantization

- ✓ Identify blocks: Structure-Aware.
- ✓ Quantize Conv: Uniform Based.
- ✓ Quantize Transformer: Log_2 Based.



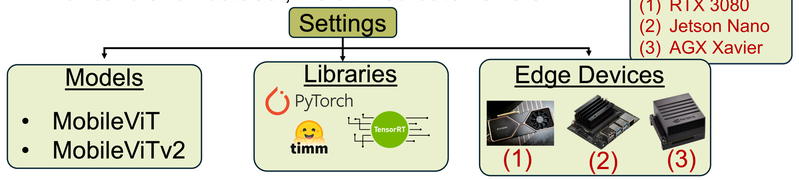
Structure-Aware Block Identification

Input: Model M
Output: Lists $CNN, Transformer$
Initialize Q, lists $CNN, Transformer$;
Enqueue $(M, "")$ into Q;
while Q *is not empty* **do**
  / Dequeue the last element from the queue */*
  $(module, parent\_name) \leftarrow$ dequeue Q;
  foreach $(layer\_name, layer)$ *in* $module.children()$ **do**
    $fullname \leftarrow parent\_name + layer\_name$;
    if *layer is* `Conv2d` **then**
      Append $fullname$ to CNN;
    else if *layer is* `Linear`*,"attn","Attention", or "Transformer"* **then**
      Append $fullname$ to $Transformer$;
    Enqueue $(layer, fullname)$ into Q;
return $CNN, Transformer$;

Experimental Settings

➢ Experiment Setup

- ❖ We used pretrained models from Timm library.
- ❖ For calibration dataset, we utilized batch size 32.



➢ Metrics

- ❖ Accuracy (Top-1)
- ❖ Latency
- ❖ Memory

➢ Existing PTQ Methods

- ❖ CNN-based
  - EasyQuant
- ❖ Transformer-based
  - PTQ4ViT
  - RepQ-ViT
- ❖ Hybrid model
  - Q-HyViT
  - HyQ

Results

➢ Accuracy

Top-1(%) accuracy of EfficientQuant with PTQ methods on the ImageNet-1K dataset.

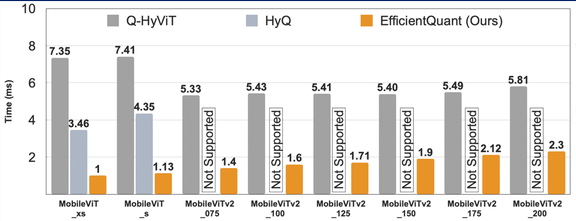
Models	FP32 (%)	PTQ for Pure CNN/ ViTs (%)			PTQ For Hybrid Models (%)				
		EasyQuant	PTQ4ViT	RepQ-ViT	Q-HyViT	Q-HyViT (Reproduced)	HyQ	HyQ (Reproduced)	EfficientQuant (Ours)
MobileViT_xxs	68.91	36.13	37.75	1.85	68.20	62.73	68.15	67.94	66.86
MobileViT_xs	74.62	73.16	65.52	41.96	74.31	69.37	73.99	73.99	73.8
MobileViT_s	78.31	74.21	68.19	59.01	77.92	75.71	77.93	77.93	77.87
MobileViTv2_050	70.1	66.80	39.39	26.60	69.89	64.99	69.16	–	65.59
MobileViTv2_075	75.57	62.91	65.54	55.52	75.29	69.81	74.47	–	72.50
MobileViTv2_100	77.9	69.34	51.02	40.61	77.63	69.66	-	–	72.03
MobileViTv2_125	79.64	77.31	67.39	41.65	79.31	71.72	-	–	77.58
MobileViTv2_150	80.36	75.83	68.61	62.12	79.97	73.09	-	–	79.08
MobileViTv2_175	80.86	79.93	72.30	63.52	80.45	75.20	-	–	80.30
MobileViTv2_200	81.12	80.04	75.50	64.65	80.76	75.85	-	-	80.48

Observation 1: HyQ achieves the highest top-1 accuracy. However, it lacks support complex MobileViTv2 variants.

Observation 2: EfficientQuant provides stable accuracy across all MobileViT models, with minimal accuracy degradation on larger variants.

➢ Latency

Observation 3: EfficientQuant outperforms Q-HyViT and HyQ in terms of latency on RTX 3080 across all MobileViT variants. Unlike HyQ, which is limited to smaller models, EfficientQuant scales effectively across the MobileViTv2 variants.

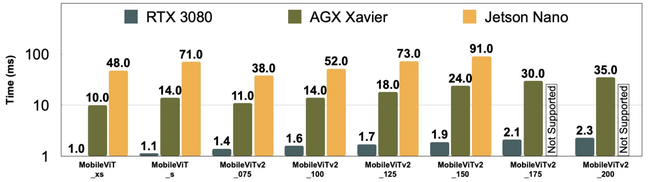


✓ EfficientQuant is 3.85× faster inference than Q-HyViT.

✓ EfficientQuant achieves **2.5× to 8.7×** reduction in latency compared to Q-HyViT

➢ Evaluation with Edge Devices

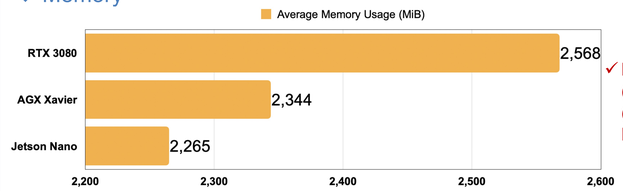
❖ Latency



✓ Complex MobileViT variants exceed memory limits

Observation 4: Jetson Nano exhibits **3×–5×** higher latency than AGX Xavier and **up to 63×** slower than RTX 3080.

❖ Memory



✓ Memory profiling: *pynvml* (RTX 3080), *tegrastats* (AGX Xavier, Jetson Nano)

Conclusion

- ✓ **Structure-aware PTQ:** Automatically identifies blocks; applies uniform (CNN) and log_2 (Transformer) quantization.
- ✓ **Performance:** Up to **8.7× faster inference** with minimal accuracy loss.
- ✓ **Edge devices:** Runs efficiently on **RTX 3080, AGX Xavier, and Jetson Nano**.

Acknowledgment

This work was supported in part by the U.S. National Science Foundation under Grant No. 2245729, and by the University Research Committee Faculty Research Fellowship at Oakland University.