When Less is More: Surprising Gains from Label-Aware Quantization

Authors

Ari Juljulian — ajuljulian@ucsd.edu
Jessica Hung — yuhung@ucsd.edu
Saathvik Dirisala — sdirisala@ucsd.edu
Yijun Luo — yil176@ucsd.edu

Mentors

Alex Cloninger — acloninger@ucsd.edu
Rayan Saab — rsaab@ucsd.edu

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

1. INTRODUCTION

What is Label-Aware Quantization (LAQ)?

- **PTQ** techniques utilize data with **similar distribution as train data** to achieve high performance under memory constraints.[1]
- **LAQ** uses data with **different distribution from train data**

Why LAQ?

- Many ML tasks involve only **subsets** of much larger datasets
- **LAQ** will reduce model size in memory, and
- **LAQ** might perform better than the original model on **subsets**

Research Question:
Does LAQ boost CNN performance on subset classification tasks?

2. DATASET

CIFAR-100:

- 100 classes of 3-channel 32x32 images
- 50,000 train; 10,000 test

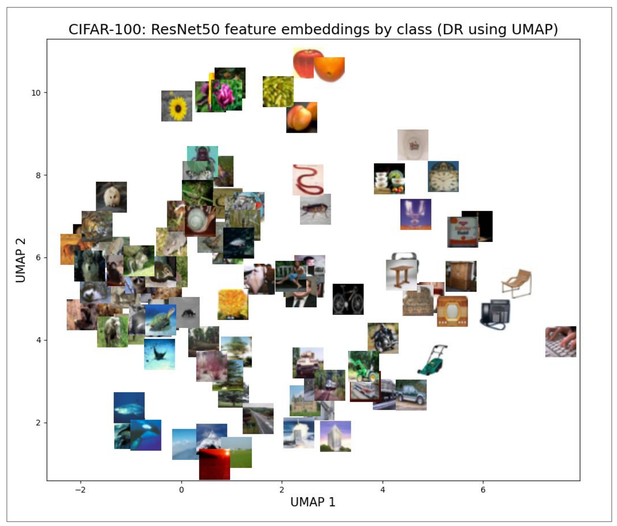


Figure 1: Visualizing Class Centers from Section 3.2

3. SUBSET GENERATION

1. **Feature Extraction (FE):** Flattened output of Conv layers of pretrained ResNet-50 (2048 dim.)
2. **Dimensionality Reduction (DR):** UMAP preserves cluster structure & location (2 dim.)
3. **Inter-Class Distance (ICD):** KL divergence[3] (Gaussian closed-form) for every unique pair of classes
4. **Subsets (SG):** Select 10 classes greedily based on an inter-class similarity parameter for spread along x-axis:
   a. **Similar Classes:** Low Median Distance
   b. **Dissimilar Classes:** High Median Distance
   c. **Random Classes:** Intermediate Median Distance

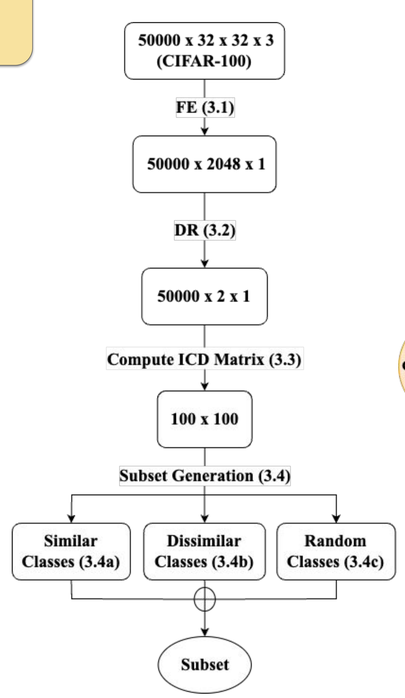


Figure 2: Data Preprocessing & Generation

4. METHODOLOGY

1. **Load Pre-Trained CNN:** CNN trained on the full dataset. Examples: ResNet, VGG, MobileNet, etc.
2. **Generate Subset (Section 3):** Used 10-class subsets
3. **Model Variations:**
   a. **Original:** Original weights
   b. **Quant:** Quantized using train split of subset
   c. **FT:** Fine-tuned using train split of subset
   d. **FT + Quant:** Fine-tuned and then quantized
4. **Evaluation (Top-1 Accuracy):**
   a. **All:** Model allowed to select from all classes
   b. **Sub:** Model allowed to select only from subset

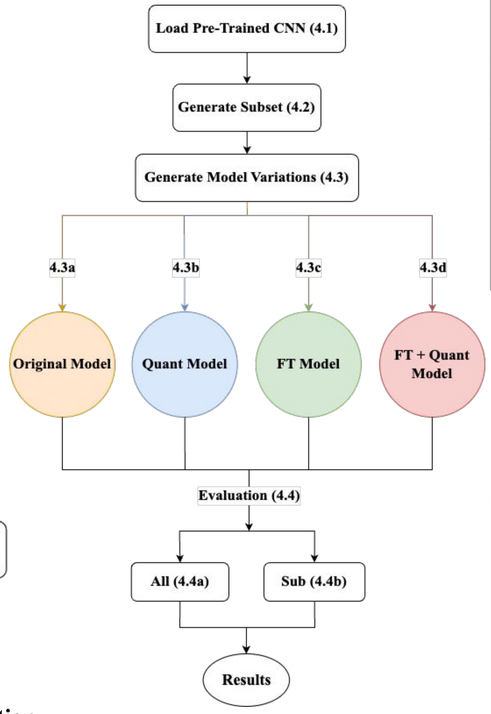


Figure 3: Experimental Setup

5. RESULTS

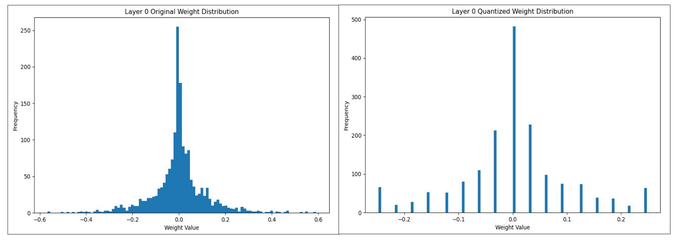


Figure 4: Impact of quantization on weight distribution

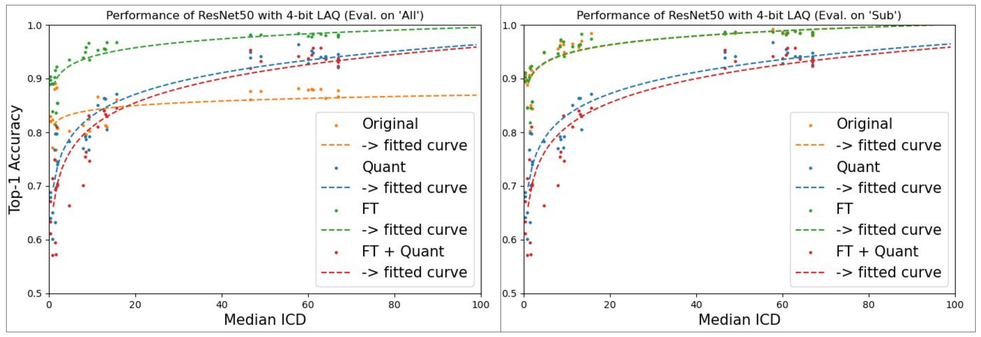


Figure 5: ResNet50 4-bit LAQ

- FT model outperforms other models (**Figure 5.1**)
- Original & FT models have identical curves (**Figure 5.2**).
- FT + Quant is more detrimental than just Quant (**Figures 5.1 & 5.2**).
- Original outperforms Quant when median ICD is low (**Figure 5.1**).

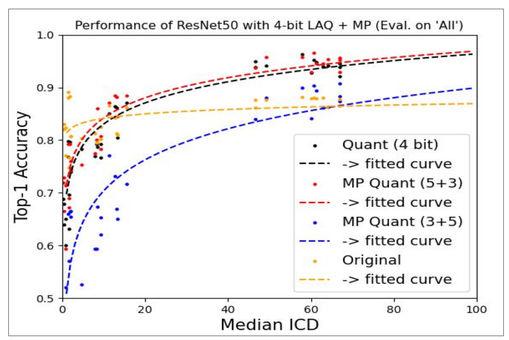


Figure 6: Mixed Precision (MP)
Maintains avg bits/wt of ~4 bits
Two variations:

1. **(5+3)**: First 50% weights = 5 bits; Second 50% weights = 3 bits
2. **(3+5)**: First 50% weights = 3 bits; Second 50% weights = 5 bits

- MP Quant (5+3) model performs slightly better than the regular 4-bit Quant model
- MP Quant (5+3) performs significantly better than MP Quant (3+5)

GPFQ (Greedy Path-Following Quantization)[2]

- Computationally efficient quantization method for pre-trained models (MLPs and CNNs).
- Quantizes each neuron using a greedy path-following algorithm, eliminating need for complex retraining.

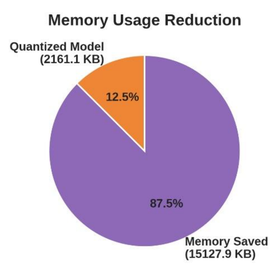


Figure 7: 4-bit LAQ Memory Footprint Reduction

Conclusions:

- LAQ boosts CNN performance only for subsets with high median ICD.
- LAQ likely performs some Fine-Tuning on CNNs.
- MP + LAQ: Earlier layers in CNNS are more sensitive to quantization

References:

[1] H. Yu et al. "Is In-Domain Data Really Needed? A Pilot Study on Cross-Domain Calibration for Network Quantization." 2021
[2] J. Zhang et al. "Post-Training Quantization For Neural Networks With Provable Guarantees." 2022
[3] S. Kullback et al. "On information and sufficiency." *The Annals of Mathematical Statistics*, 22(1), 79–86. 1951.

<INSERT QR CODE>

When Less is More: Surprising Gains from Label-Aware Quantization

Ari Juljulian ajuljulian@ucsd.edu
Jessica Hung yuhung@ucsd.edu
Saathvik Dirisala sdirisala@ucsd.edu
Yijun Luo yil176@ucsd.edu
Alex Cloninger acloninger@ucsd.edu
Rayan Saab rsaab@ucsd.edu

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

1. INTRODUCTION

What is Label-Aware Quantization (LAQ)?

- **Quantization** techniques aim to strategically lower bit width of weights while minimizing loss of a **wide range** outputs.
- **LAQ** focuses on a **narrower range** of outputs/classes.

Why LAQ?

- Many ML tasks involve only **subsets** of a larger dataset
- **LAQ** will reduce model size in memory, and
- **LAQ** might perform better than the original model on **subsets**

Research Question:
Does LAQ boost CNN performance for subsets?

2. DATASET

CIFAR-100:

- 100 classes of 3-channel 32x32 images
- 50,000 train; 10,000 test

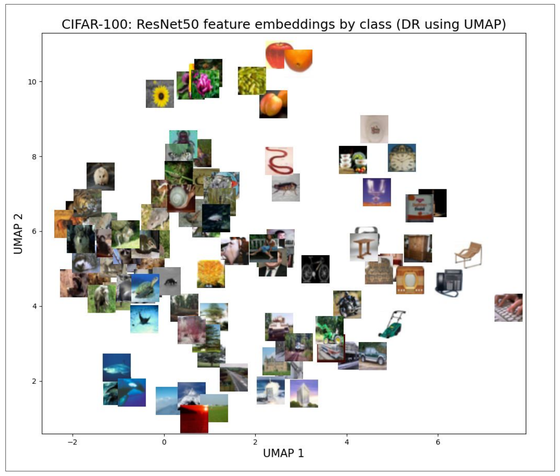


Figure 1: Visualizing Class Centers from Section 3.2

3. SUBSET GENERATION

1. **Feature Extraction (FE):** Flattened output of Conv layers of pretrained (ImageNet) ResNet-50 (2048 dim.)
2. **Dimensionality Reduction (DR):** UMAP preserves cluster structure & location (2 dim.)
3. **Inter-Class Distance (ICD):** KL divergence (Gaussian approx.) for every unique pair of classes
4. **Subsets (SG):** Select 10 classes greedily based on an inter-class similarity parameter for even spread along x-axis:
   a. **Similar Classes:** Low Median Distance
   b. **Dissimilar Classes:** High Median Distance
   c. **Random Classes:** Intermediate Median Distance

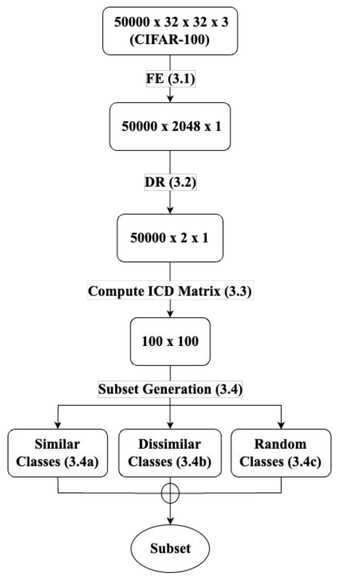


Figure 2: Data Preprocessing & Generation

GPFQ (Greedy Path-Following Quantization):

- Computationally efficient quantization method for pre-trained models (MLPs and CNNs).
- Quantizes each neuron using a greedy path-following algorithm, eliminating need for complex retraining.

4. METHODOLOGY

1. **Load Pre-Trained CNN:** CNN trained on the full dataset. Examples: ResNet, VGG, AlexNet, MobileNet, etc.
2. **Generate Subset (Section 3):** Used 10-class subsets
3. **Model Variations:**
   a. **Original:** Original weights
   b. **Quant:** Quantized using train split of subset
   c. **FT:** Fine-tuned using train split of subset
   d. **FT + Quant:** Fine-tuned and then quantized
4. **Evaluation (Top-1 Accuracy):**
   a. **All:** Model allowed to select from all classes
   b. **Sub:** Model allowed to select only from subset classes

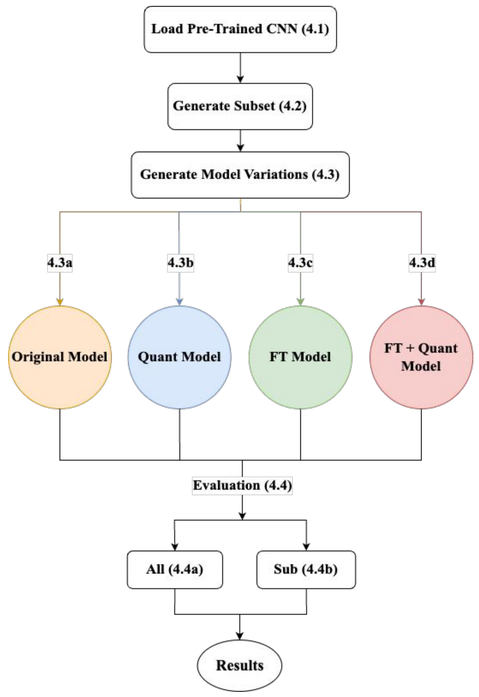
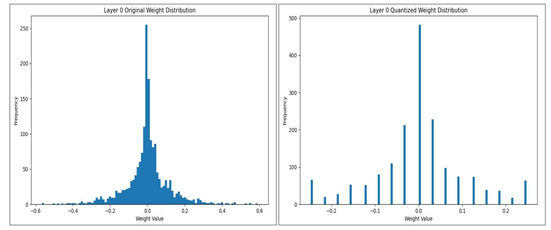
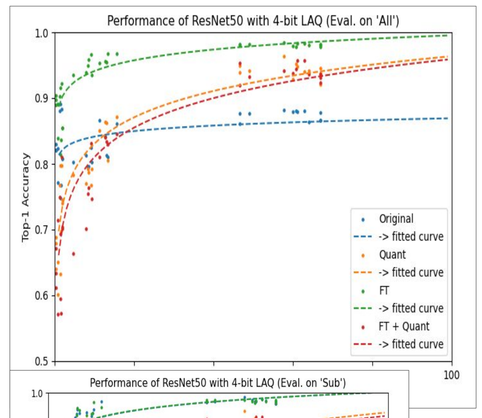
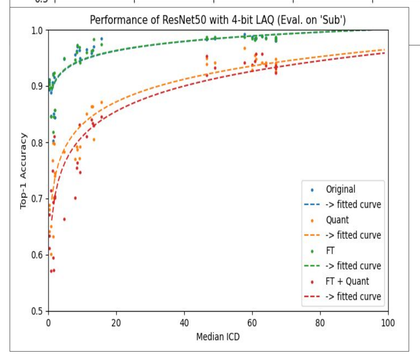


Figure 3: Experimental Setup

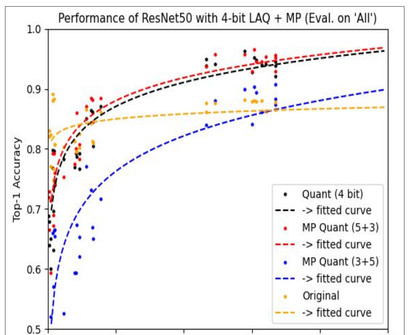
5. RESULTS

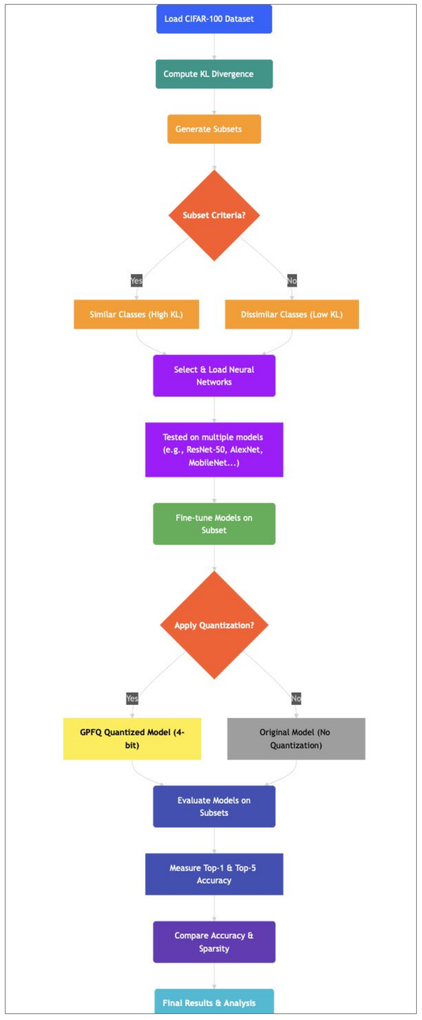






- Original & Fine-Tuned models have identical curves (**Figure 3**); Fine-Tuned model outperforms other models (**Figure 4**).
- Quantization after Fine-Tuning is more detrimental than just Quantization (**Figures 3 & 4**).
- Original model outperforms Quantized models when median inter-class distance is low (**Figure 4**).





5. RESULTS

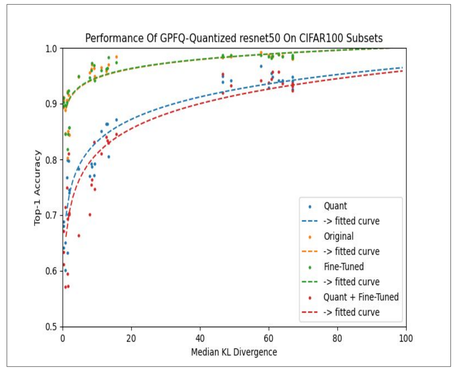


Figure 3: Output restricted to only subset classes

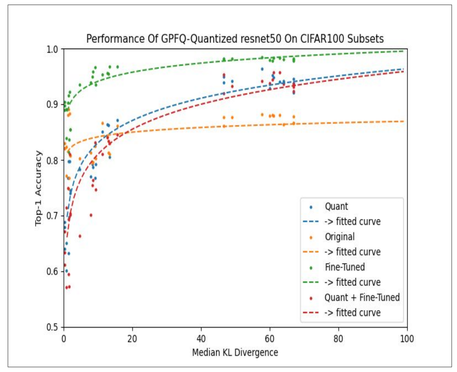


Figure 4: No restriction

Parameters:

- Model: ResNet-50
- Bit Width: 4

- Original & Fine-Tuned models have identical curves (**Figure 3**); Fine-Tuned model outperforms other models (**Figure 4**).
- Quantization after Fine-Tuning is more detrimental than just Quantization (**Figures 3 & 4**).
- Original model outperforms Quantized models when median inter-class distance is low (**Figure 4**).

Mixed Precision

- Maintains average bits/weight = ~4 bits
- 2 variations:
  1. **(5+3)**: First 50% = 5 bits; Second 50% = 3 bits
  2. **(3+5)**: First 50% = 3 bits; Second 50% = 5 bits
- Result measured with no restriction

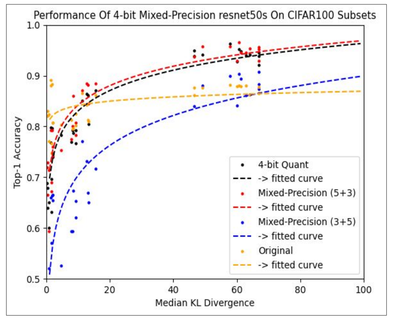


Figure 5: Mixed-precision label-aware quantization

6. CONCLUSIONS

1. Fine-tuning a pretrained CNN on a subset of the original dataset yields the best results.
2. Quantizing a pretrained CNN using a subset of the original dataset lowers bit width and also fine-tunes the outputs.
3. When no restriction is placed on the output, Quantized models continue to classify with high accuracy. This may imply that CNN knowledge associated with classes not in subset is compromised to maintain subset accuracy in the process of post-training quantization.