DSC180 Q2 Project Report

Saathvik Dirisala
sdirisala@ucsd.edu

Jessica Hung
yuhung@ucsd.edu

Yijun Luo
yil176@ucsd.edu

Ari Juljulian
ajuljulian@ucsd.edu

Rayan Saab
rsaab@ucsd.edu

Alex Cloninger
acloninger@ucsd.edu

Abstract

Image classification models are commonly benchmarked on the ImageNet dataset, which contains approximately 1,000 object classes and rigorously tests model performance. Convolutional Neural Networks (CNNs), the revolutionary architecture for image classification, have increased in size to meet the growing complexity of tasks. To enable deployment on edge devices with limited memory and computational power, various quantization techniques have been developed. Among these, the Greedy Path-Following Quantization (GPFQ) algorithm achieves 4-bit and 3-bit precisions with minimal accuracy loss on ImageNet. This study explores label-aware quantization, with the aim of analyzing the quantizability of CNNs when the task is restricted to subsets of data, in particular the CIFAR100 dataset. We empirically analyze the impact of sub-task calibration data on classification accuracy, investigating in-depth the relationship between subset similarities to performance. Our findings provide insights into optimizing CNN quantization for sub-task specific applications.

1 Introduction

As neural networks become increasingly integral to modern life as part of AI, their environmental and computational costs increase significantly. Training and deploying them often require vast energy resources, posing sustainability concerns. Moreover, many state-of-the-art models, such as convolutional neural networks (CNNs), are too large for practical deployment on edge devices such as smartphones or IoT sensors.

Quantization, the process of reducing the precision of model parameters, is a key solution to this challenge. By shrinking the size of the models, quantization enables faster inference and lower energy consumption while maintaining performance. However, existing methods typically treat all tasks equally, ignoring opportunities to specialize compression for specific subsets of data or tasks. Our project seeks to address this gap by exploring label-aware and mixed-precision quantization to enhance the deployability of CNNs for specific applications. This effort not only aligns with the growing demand for efficient AI but also has the potential to reduce the carbon footprint of machine learning systems.

The goal of this project is to integrate label-aware quantization and mixed-precision quantization into a unified framework for optimizing CNN compression. Specifically: Label-Aware Quantization.

The prior literature on GPFQ quantization shows the underlying mathematical proof of error bounds of such model compression based on data distribution. The difference between post-training quantization (PTQ) and quantization-aware training (QAT). PTQ involves first training the model, then the range of representative data input taken by each parameter is dissolved into bins for the mapping of the original values. Without the need to further train the model, PTQ is generally more efficient than QAT. QAT, on the other hand, is a fine-tuning process of a model where the model is further trained with the objective of quantization. With GPTQ, a more time-efficient method of quantizing weights is introduced.

Neural networks are generally pre-trained on datasets with hundreds or thousands of classes. However, in application, we may only be interested in a subset of the classes. With this notion in mind, we wish to only quantize the entire network with the data labeled with the classes of interest. Label-aware quantization may be able to enhance the accuracy performance and efficiency of the model than the original model. We will further explore how the similarities between classes of data subsets may affect the quantizability and performance of the model.

2 Methodology

In this study, we employed a ResNet50 architecture to perform experiments on the CIFAR-100 dataset. The dataset was modified to include only a subset of its original classes, specifically focusing on various 10-class subsets drawn randomly from the full dataset. The goal was to analyze the effects of quantization on model performance and inter-class distances within these subsets.

2.1 Model and Dataset Preparation

The ResNet50 model was quantized down to 4 bits using the Greedy Path-Following Quantization (GPFQ) algorithm. This quantization technique aims to reduce the model size and computational requirements while maintaining a competitive level of accuracy.

For each experiment, we selected different 10-class subsets from CIFAR-100. These subsets were used to train and evaluate the quantized model. The median inter-class distances were recorded for each subset to understand how class separation influences model performance post-quantization.

We hypothesized that there would be a strong correlation between average KL divergence and accuracy. If the selected subset contains classes that are very different from each other, it would be easier to differentiate between them, resulting in higher accuracy. The opposite would also be true: subsets containing similar classes would yield lower accuracy as it would be more difficult to distinguish between them. In order to quantify the similarity between classes, the Kullback-Leibler (KL) divergence was used. A subset generation was implemented which generated highly similar, highly dissimilar, or random subsets. The function would begin by selecting a random class and iteratively add classes depending on the average KL divergence with the already existing classes. By doing this, we can ensure that there is an effective method to systematically generate highly similar or highly dissimilar class subsets.

2.2 Inter-Class Distance Computation

The inter-class distance was calculated using the following procedure:

1. **Embedding Extraction:** We retrieved the flattened embeddings from the final convolutional layer of the ImageNet-pretrained ResNet50 model for every training data item within the CIFAR-100 dataset.
2. **Dimensionality Reduction:** The high-dimensional embeddings were reduced to two dimensions using Uniform Manifold Approximation and Projection (UMAP) (1). As opposed to other dimensionality reduction techniques such as t-SNE, UMAP is better able to maintain both local and global features without requiring much computational power. By reducing the embeddings to 2D, we were able to visualize class separability and measure distances between the classes. This step helped in simplifying the complexity of the data while preserving the underlying structure necessary for distance computation.
3. **KL Divergence Calculation:** The 2D distribution of the embedding representations for each class was used to compute the Kullback-Leibler (KL) divergence score between every pair of classes. Given that KL divergence is not symmetric, we calculated the divergence in both directions for each class pair and then averaged these values to obtain a symmetric measure of inter-class distance. The KL divergence scores were computed under the assumption that the 2D distributions follow a Gaussian distribution.

4. **Distance Matrix Construction:** Using the computed KL divergence scores, we constructed a 100x100 distance matrix, representing the inter-class distances for all classes within the CIFAR-100 dataset. All diagonal entries would be 0s as that would be the distance of a class to itself. So, the minimum KL divergence two classes can have would be 0 while the maximum would be $+\infty$. However, the maximum is ultimately dependent on how distinct the classes are in the 2D UMAP space.

2.3 Performance Evaluation

To evaluate the impact of quantization, we generated a Quantized Curve using the inter-class distance data. We compared the performance of the quantized ResNet50 model against two baselines:

- **Original Model Accuracy:** The accuracy of the original, non-quantized ResNet50 model pretrained on CIFAR-100 was measured on each of the 10-class subsets.
- **Fine-Tuned Model Performance:** The pretrained ResNet50 model was fine-tuned on each of the 10-class subsets. Post fine-tuning, we evaluated the model's accuracy to understand the benefits of additional training on the subsets.

By comparing these results, we aim to assess the trade-offs between quantization-induced efficiency gains and potential accuracy degradation, as well as the role of inter-class distances in these dynamics.

In future experiments, we plan on gathering experimental data for other bit-widths (ex: 3, 8, 16) and model architectures (ex: VGG16, DenseNet121, Vision Transformer).

3 Results

In this section, we present the performance outcomes of the quantized ResNet50 model compared to the original and fine-tuned models on various 10-class subsets of the CIFAR-100 dataset.

3.1 Model Performance

The accuracy of the original, non-quantized ResNet50 model, the quantized model (using the GPFQ algorithm), and the fine-tuned model were evaluated on each of the 10-class subsets. The quantized model exhibited varying levels of accuracy depending on the inter-class distances within each subset.

Observations from Figure 3

- **Original Model Accuracy:** The original model maintained higher accuracy across most subsets, serving as the baseline for comparison.

- **Quantized Model Performance:** A noticeable drop in accuracy was observed post-quantization, which varied based on the inter-class distances.
- **Fine-Tuned Model Accuracy:** Fine-tuning the model on the specific subsets improved performance compared to the quantized model, though it did not always match the original model's accuracy.

3.2 Inter-Class Distance Analysis

	Class Pair	Avg KL Divergence
0	(boy, girl)	0.038
1	(bee, butterfly)	0.060
2	(girl, woman)	0.061
3	(otter, seal)	0.074
4	(crab, lobster)	0.101

Figure 1: Some of the closest CIFAR-100 classes based on our distance metric

In Figure 1, we can see some of the closest classes among all possible pairings in CIFAR-100. It makes intuitive sense that these are some of the closest in distance.

In Figure 2, we see that the inter-class distance distribution for each subset is highly skewed. Given this, it makes sense to use th emedian as the central tendency to represent the plot.

The inter-class distance matrix, constructed using KL divergence, revealed a correlation between class separation and model accuracy. Subsets with larger median inter-class distances generally exhibited higher accuracy, particularly in the quantized model.

Figure 3 illustrates the performance trends across different subsets, highlighting the relationship between inter-class distances and model accuracy.

4 References

References

[1] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. Accessed via online article.

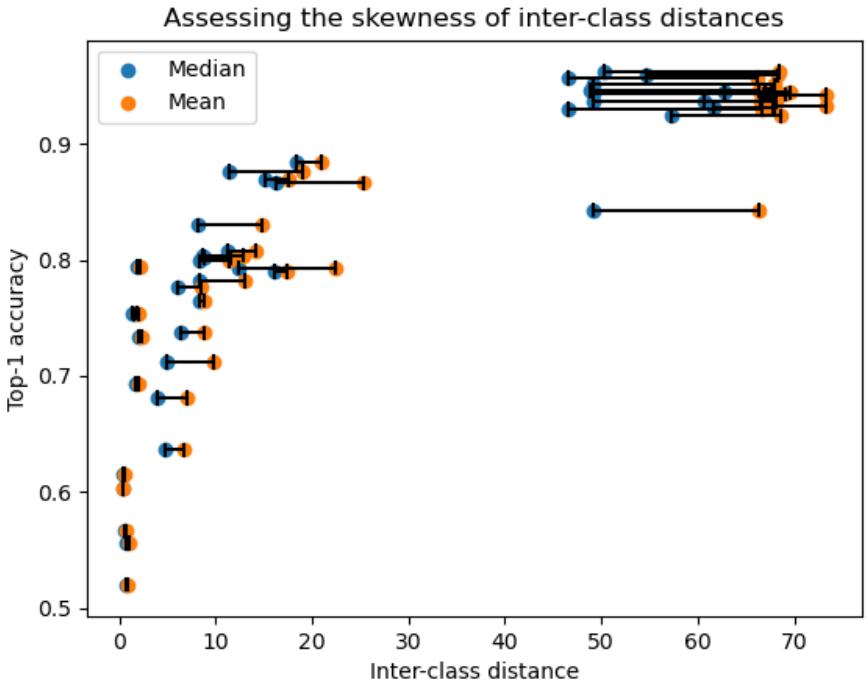

Figure 2: Assessing the skewness of inter-class distances

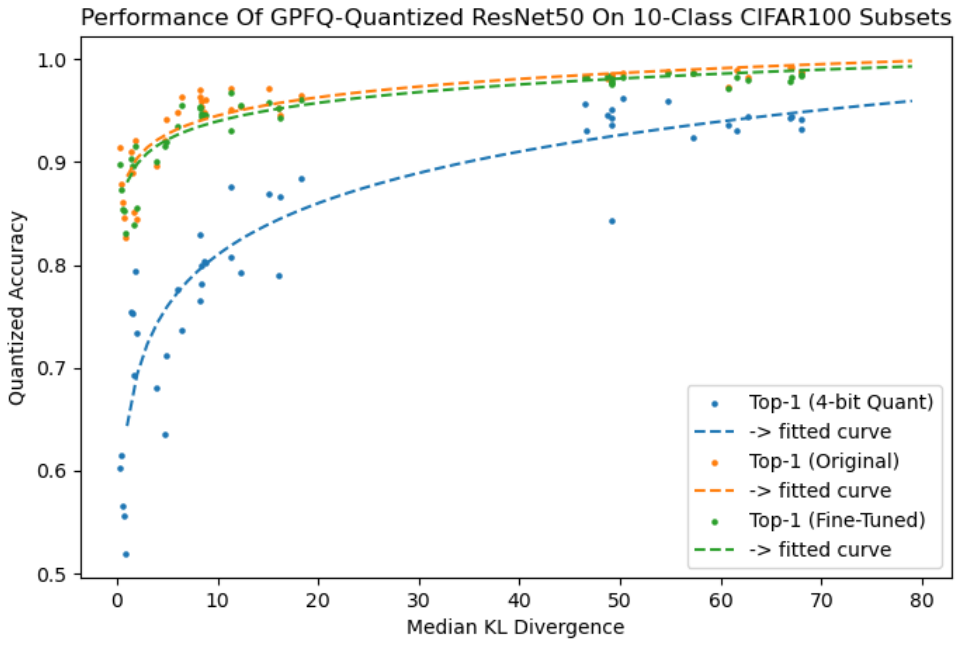

Figure 3: Performance comparison for 4-bit quantized ResNet50 using various subsets

5 Appendix

5.1 Abstract

Image classification models are commonly benchmarked on the ImageNet dataset, which contains approximately 1,000 object classes and rigorously tests model performance. Convolutional Neural Networks (CNNs), the state-of-the-art architecture for image classification, have been increasing in size to meet the growing complexity of tasks. To enable deployment on edge devices with limited memory and computational power, various quantization techniques have been developed. Among these, the Greedy Path-Following Quantization (GPFQ) algorithm achieves 4-bit and 3-bit precisions with minimal accuracy loss on ImageNet. This study explores label-aware quantization, aiming to analyze the quantizability of CNNs when the task is restricted to subsets of ImageNet. We empirically analyze the impact of sub-task calibration data on classification accuracy, compare techniques like GPFQ and BRECQ, investigate mixed-precision quantization, and extend the GPTQ algorithm—originally designed for LLM compression—to CNN architectures. Our findings provide insights into optimizing CNN quantization for sub-task-specific applications.

5.2 Broad Problem Statement

As neural networks become increasingly integral to modern life as part of AI, its environmental and computational costs grow significantly. Training and deploying them often require vast energy resources, posing sustainability concerns. Moreover, many state-of-the-art models, such as convolutional neural networks (CNNs), are too large for practical deployment on edge devices like smartphones or IoT sensors.

Quantization, the process of reducing the precision of model parameters, is a key solution to this challenge. By shrinking model sizes, quantization enables faster inference and lower energy consumption while maintaining performance. However, existing methods typically treat all tasks equally, ignoring opportunities to specialize compression for specific subsets of data or tasks. Our project seeks to address this gap by exploring label-aware and mixed-precision quantization to enhance the deployability of CNNs for specific applications. This effort not only aligns with the growing demand for efficient AI but also has the potential to reduce the carbon footprint of machine learning systems.

5.3 Narrow Problem Statement

The goal of this project is to integrate label-aware quantization and mixed-precision quantization into a unified framework for optimizing CNN compression. Specifically:

5.4 Label-Aware Quantization

The prior literature on GPFQ quantization shows the underlying mathematical proof of error bounds of such model compression based on data distribution. The difference between post-training quantization (PTQ) and quantization-aware training (QAT). PTQ involves first training the model, then the range of representative data input taken by each parameter is dissolved into bins for the mapping of the original values. Without the need to further train the model, PTQ is generally more efficient than QAT. QAT, on the other hand, is a fine-tuning process of a model where the model is further trained with the objective of quantization. With GPTQ, a more time-efficient method of quantizing weights is introduced.

Neural networks are generally pretrained on datasets with hundreds or thousands of classes. However, in application, we may only be interested in a subset of the classes. With this notion in mind, we wish to only quantize the entire network with the data labeled with the classes of interest. Label-aware quantization may be able to enhance the accuracy performance and efficiency of the model than the original model. We will further explore how the distribution of data subsets may affect the quantizability and performance of the model. For example, if the distribution of the subclasses follow a similar form as the distribution of the entire dataset on which the original model is trained on, the performance in terms of efficiency and accuracy may be higher. However, if the distributions are disimilar significantly, the quantizability may be negatively affected.

5.5 Adapting GPTQ for CNN Quantization

GPTQ is a state-of-the-art quantization algorithm that is popularly used for compressing LLMs. This method employs second-order information on weights in a specific layer to re-distribute quantization error. It also implements a parallelized approach to the quantization of weights. These two techniques combine to effectively compress LLMs down to 4-bit precision. However, given the effectiveness of GPTQ in the realm of LLMs, it can logically be posited that a similarly-structured algorithm tailored for CNN architectures could potentially bring about high levels of compression in CNN models. BRECQ is one algorithm that is effective in exploiting second-order information to quantize CNN models. However, the algorithm has two downsides: it uses an iterative approach to quantize the weights, and it quantizes both weights and activations. The former feature mentioned increases the computational cost for quantizing the network, and latter excessively depletes model performance. GPTQ, on the other hand, is a single-shot quantization algorithm that approximates the Hessian (rather than computing it with high precision) and only quantizes weights. Therefore, GPTQ reduces the algorithmic computational cost, in relation to BRECQ. We wish to then see how GPTQ compression of CNNs compares to GPFQ and BRECQ counterparts. Observing performance on sub-tasks of ImageNet can help us effectively gather results regardless of low compute resources.

5.6 Mixed-Precision Quantization

As neural networks have become increasingly complex, their extensive computational demands make it difficult to use them on personal devices like smartphones and sensors. To reduce this computational burden, quantization techniques are used where the precision of the network is significantly reduced while accuracy is maintained. For example, a 32-bit model can be reduced to an 8-bit model, while the dropoff in accuracy is minimal. A quantization method that I find particularly interesting is mixed precision quantization. Rather than uniformly quantizing the model to where all of the layers have the same precision level, we can fine-tune the model where more important layers are allowed higher precision while less important layers are more heavily quantized. This will be performed on various common neural networks such as ResNet and AlexNet using datasets such as ImageNet and CIFAR-10. There are various methods that we can explore that will help us understand the architecture of neural networks and identify which layers are more important. For example, we can keep track of the current layer, the previous layer, and the next layer. If we see that a shrinking process is taking place, we can potentially make the assumption that these layers contribute less to the overall performance of the model. Through techniques such as this, mixed precision quantization can be an extremely effective method of quantization.

5.7 Primary Output and Success Justification

In our project, we aim to combine the principles of label-aware quantization and mixed-precision quantization to create a unified framework for optimizing neural network compression. Label-aware quantization focuses on leveraging task-specific sub datasets for improved quantizability, while mixed-precision quantization fine-tunes model precision at a layer-specific level based on their contribution to overall network performance. By integrating these two approaches, we hypothesize that it is possible to achieve a higher degree of compression efficiency without sacrificing accuracy, particularly for sub-task-specific datasets.

The key idea is to first analyze the data distribution of the targeted subset of labels and identify the layers most critical to preserving task-specific features. With this insight, we will apply mixed-precision quantization, prioritizing higher precision for layers contributing significantly to the classification accuracy of the selected labels.

To examine how particular classes of data affects the quantization sensitivity of a CNN, we will adopt the manifold approximation for visualization.

The primary output of this project will be a paper detailing our findings and methodologies. It will include: Empirical results comparing GPFQ, BRECQ, and GPTQ applied to CNNs; Visualization of data distributions and layer importance for mixed-precision quantization; Recommendations for deploying quantized CNNs in task-specific applications.

Since we will utilize the ImageNet dataset already processed in the Quarter 1 Project, data availability and quality are assured. The subset classes will be selected based on relevance

to target tasks, ensuring meaningful analyses. Our project builds on robust existing frameworks, which increases the likelihood of success.

6 Contributions

6.1 Saathvik

- Altered the GPFQ repository to work for CIFAR-100 subsets and for a pre-trained ResNet50.
- Developed the function to generate an inter-class similarity score, researching UMAP for dimensionality reduction and CNN embeddings for feature extraction.

6.2 Jessica

- Revised the GPFQ script to enable experiments on the CIFAR-100 dataset using pre-trained ResNet20 and MobileNetV2_x0_5.
- Compared accuracy results across three scenarios: the original model, the model quantized using the entire training dataset, and the model quantized using only a specific data subset of interest.
- Visualized final accuracies across models on more than 50 randomly generated 10-class subsets to analyze performance differences.

6.3 Ari

- Ran experiments for 3-bit, 4-bit, and 8-bit quantizations and plotted the results
- Wrote function to generate subsets with equal distance from each other by passing in the following parameters: min divergence, max divergence, and range
- Replicated results of the ResNet-50 model with AlexNet in order to identify patterns with the quantization methods

6.4 Yijun

- Researched multiple metrics for evaluating classes distance.
- Experimented and replicated subset quantization with various model architectures, including VGG16, EfficientNet, and DenseNet121.