Balancing Accuracy and Efficiency: A Comparative Study of Knowledge Distillation and Post-Training Quantization Sequences

Daniel Gitelman dgitelman@ucsd.edu

Jiangqi Wu jiw118@ucsd.edu

Vishwak Pabba vpabba@ucsd.edu

Zhiqing Wang zhw055@ucsd.edu

Alex Cloninger acloninger@ucsd.edu

Rayan Saab rsaab@ucsd.edu

Abstract

This study examines the interplay between knowledge distillation and post-training quantization techniques to optimize deep neural networks (DNNs) for deployment in resource-constrained environments. We systematically evaluate various distillation methods—including Vanilla Knowledge Distillation (VKD), Mixup augmentation, Deep Mutual Learning (DML), and Decoupled Knowledge Distillation (DKD)—applied to compressing ResNet50 models into smaller ResNet18 models. Post-training quantization using a greedy path-following algorithm further reduces model size and computational load. Experimental results on CIFAR-10 and CIFAR-100 datasets demonstrate that, while distilled student models effectively retain accuracy at moderate quantization levels (around 8-bit), accuracy sharply declines at lower bit widths, especially on complex datasets like CIFAR-100. This marks student model are less compressible at lower bits, the model compression strategy for combining knowledge distillation with quantization is still significant with highly-constrained deployment environment and dealing with less complex tasks.

Website: https://sicily496.github.io/DSC180B_Q2_Project_website/
Code: https://github.com/jiangqiw/DSC180B_Q2_Project

1 Introduction

Deep neural networks (DNNs) have become the cornerstone of modern machine learning, achieving unprecedented performance in critical tasks such as medical image diagnosis, autonomous driving, and real-time language translation. For instance, architectures like ResNet50 achieve over 92% top-1 accuracy on benchmark datasets like CIFAR-10, rivaling human-level performance in specific domains. However, these models often demand extensive computational resources: ResNet50, for example, comprises 25.6 million parameters and requires approximately 100 MB of storage, making deployment on resource-constrained platforms—such as mobile devices, IoT sensors, or edge computing systems—prohibitively challenging. To address this gap, model compression techniques have emerged as a vital area of research, aiming to reduce computational and storage overhead while preserving accuracy. Existing methods, such as knowledge distillation (transferring knowledge from a large ”teacher” model to a compact ”student” model) and quantization (reducing numerical precision of weights), have shown promise individually. However, the interplay between these techniques—particularly their combined impact on accuracy and storage efficiency—remains underexplored. In this project, we propose a hybrid compression strategy that systematically integrates knowledge distillation and post-training quantization to optimize the trade-off between model efficiency and performance. Specifically, we focus on compressing ResNet50. Our methodology involves two stages, first involved in implementing various distillation strategies to facilitate knowledge transfer from ResNet50 to ResNet18, followed by post-training quantization to further reduce its memory storage. By evaluating our approach on the CIFAR-10 and CIFAR-100 datasets, we demonstrate that this sequential integration of techniques achieves superior compression rates compared to single methods while maintaining competitive accuracy in less complex dataset like CIFAR-10. Conversely, for more complex datasets like CIFAR-100, ResNet18 exhibits a more significant drop in accuracy, particularly when quantized to fewer bits, such as 2 bits. Our work provides actionable insights for deploying high-accuracy DNNs in environments with strict storage and computational constraints, contributing to the broader goal of sustainable and accessible AI.

1.1 Technical Background

Model compression is a vital area of research focused on reducing the resource requirements of DNNs without substantially compromising their performance. Two prominent techniques in this domain are knowledge distillation and quantization.

Knowledge Distillation

Knowledge distillation (KD) is a technique where a smaller, compact student model learns to replicate the behavior of a larger teacher model. The student model is trained using the softened output probabilities of the teacher model, capturing the dark knowledge that encompasses the teacher’s generalization capabilities. KD helps in transferring knowledge effectively, enabling the student model to achieve performance comparable to the teacher model despite having fewer parameters.

Quantization

Quantization involves reducing the number of bits used to represent each model parameter. By converting 32-bit floating-point representations to lower bit-width formats, quantization reduces both the model size and the computational load during inference.

1.2 Prior Work

1.2.1 Knowledge Distillation

Knowledge distillation, first introduced by Bucilǎ et al. in 2006, focuses on training a smaller, more manageable student model to approximate the complex behavior of a larger teacher model through the use of pseudo data generated by ensemble models (Bucilǎ, Caruana and Niculescu-Mizil 2006). This foundational concept was further refined by Hinton et al. in 2015, who advanced the use of soft targets (or class probabilities) derived from the teacher's output to enhance the student's learning process, revealing a richer spectrum of information than traditional hard targets (Hinton, Vinyals and Dean 2015).

The application of knowledge distillation in model quantization has been explored in previous research, notably by Elthakeb et al., who demonstrated the benefits of applying knowledge distillation after quantization to enhance the performance of low-precision student models under the guidance of high-precision teacher networks (Elthakeb et al. 2020). In contrast, our study proposes to invert this sequence by initiating the process with knowledge distillation followed by quantization. This methodological shift is intended to assess whether such an approach can yield enhanced model efficiency. Additionally, our research adjusts the number and bit-width of the parameters to optimize performance and extends the investigation to various distillation techniques to evaluate their impact on the subsequent quantization process. This novel approach seeks to fill a significant gap in the existing literature, offering new insights into the synergies between knowledge distillation and model quantization for improving the deployment efficiency and performance of neural networks.

1.2.2 Quantization

Quantization, as a technique for model compression, was first introduced in the 1990s to reduce the computational and memory requirements of neural networks. The core idea behind quantization is to represent model parameters with fewer bits, typically reducing the bit-width used to store weight values from 32-bit floating point precision to lower bit-widths such as 8-bit or even binary values. By doing so, quantization significantly reduces the storage footprint of the model and accelerates inference, without sacrificing too much in terms of model accuracy. Traditional quantization methods generally apply fixed bit-widths uniformly across the entire network, simplifying the process but often resulting in a loss of accuracy due to the indiscriminate application of lower bit-widths, especially in more sensitive layers of the network.

The method of quantization that we focus on in this study is the **Greedy Path Following Quantization (GPFQ)** algorithm, a novel approach introduced by Rayan et al. (Zhang, Zhou and Saab 2023). GPFQ refines traditional quantization techniques by minimizing quantization error iteratively, all while preserving high model fidelity. Unlike conventional methods that apply uniform quantization across all layers, GPFQ adapts the quantization strategy for each layer of the network, balancing precision and performance based on the specific characteristics of each layer. This adaptive, layer-wise strategy is driven by a greedy approach that selects the most appropriate bit-width for each layer, aiming to achieve the best trade-off between compression and accuracy. Through this greedy process, GPFQ is able to maintain or even improve the model's accuracy at lower bit-widths compared to other state-of-the-art quantization methods.

Empirical evaluations have shown that GPFQ consistently outperforms existing quantization techniques, particularly in scenarios where high compression ratios are required. For example, on challenging benchmarks such as ImageNet, GPFQ achieves substantial reductions in model size while maintaining competitive or even superior accuracy. This makes GPFQ particularly valuable in resource-constrained environments, where efficient deployment of deep learning models is critical. By iterating on the quantization process and fine-tuning layer-specific precision, GPFQ offers a more flexible and effective solution for compressing neural networks without sacrificing performance, highlighting its potential for real-world applications in mobile and edge computing devices.

1.3 Current Methods Combining Quantization and Distillation

Quantization-Aware Knowledge Distillation (QKD) is a prominent method that combines the strengths of Quantization-Aware Training (QAT) and Knowledge Distillation (KD) to create efficient and compact deep learning models. Quantization-Aware Training simulates the behavior of quantized weights and activations during the training process. This allows the model to adapt to reduced precision, mitigating the performance degradation that often occurs when deploying quantized models. Knowledge Distillation, on the other hand, transfers knowledge from a high-accuracy teacher model to a smaller student model by minimizing a loss function that aligns the student's outputs with those of the teacher.

In QKD, the student model is not only compact but also explicitly trained to operate under quantization constraints. This integrated approach ensures that the student model benefits from both the adaptability of QAT and the accuracy retention capabilities of KD. Recent advancements have further refined this technique through methods like Adaptive Quantization with Distillation. Here, quantization parameters, such as bit-width, are dynamically adjusted during the distillation process based on the complexity of different model layers. This dynamic adjustment enables more efficient compression while minimizing performance trade-offs. Models trained using these integrated techniques have achieved significant reductions in size and computational requirements while maintaining competitive accuracy, making them well-suited for deployment on resource-constrained devices such as mobile phones and edge devices.

2 Experiment Methods

2.1 Experiment Methods Overview

Our methodology combines knowledge distillation (KD) and post-training quantization to compress ResNet50 into a resource-efficient ResNet18-based model. First, we systematically evaluate different KD strategies. Second, we apply a greedy path-following quantization algorithm to compress the distilled student model to lower bit sizes, iteratively selecting layers for quantization based on accuracy impact. By testing combinations of distillation strategies and quantization sequences, we determine the configuration that maximizes accuracy under strict storage constraints. Since the experiment involved in combining knowledge distillation and quantization, the following section will be divided into the knowledge distillation part and the quantization part.

2.2 Knowledge Distillation

We investigate a range of knowledge distillation techniques to discern general performance trends when combined with quantization methods. This exploration aims to facilitate a more nuanced comparison between the efficacy of combined methodologies and singular approaches relative to specific storage capacities.

2.2.1 Teacher Model

In this experiment, the teacher models employed for the CIFAR-10 and CIFAR-100 datasets are ResNet50 architectures, each pretrained and specifically customized by Eduardo Dadalto. These models are available on the Hugging Face platform.

2.2.2 Vanilla Knowledge Distillation

Vanilla Knowledge Distillation constitutes the traditional method by (Hinton, Vinyals and Dean 2015) wherein a student model is trained using a hybrid loss that integrates L_C , cross-entropy from ground-truth labels, and D_{KL} , Kullback-Leibler divergence from the teacher's soft targets, with class probabilities refined through temperature scaling T, as shown in following formula:

$$L = (1-\alpha)\cdot L_C(y,\hat{y}) + \alpha \cdot T^2 \cdot D_{KL}(p,q)$$

where α is a hyperparameter that controls the relative contribution of the cross-entropy and the KL-divergence components to the total loss, adjusting the balance between adhering to the true data distribution and mimicking the teacher's output distribution.

2.2.3 Mixup Method for Data Generation

Mixup is a data augmentation method we implemented in order to improve the overall performance of the knowledge distillation algorithm which was discussed by Beyer et al. (Beyer et al. 2022). It works by generating new training samples through the linear interpolation of pairs of images and their corresponding labels. Specifically, given two randomly selected samples (x_i, y_i) and (x_j, y_j), Mixup creates a new sample $(\tilde{x}, \tilde{y})$ using the following formulas:

$$\tilde{x} = \lambda x_i + (1 - \lambda) x_j$$
$$\tilde{y} = \lambda y_i + (1 - \lambda) y_j$$

where λ is drawn from a Beta distribution Beta(α, α). By blending both input data and their labels, Mixup encourages the model to learn smoother decision boundaries, making it more robust to adversarial examples and improving generalization. In our implementation, we applied Mixup to augment the training set before distillation, allowing the student model to learn from a more diverse and interpolated feature space, which ultimately enhanced its performance.

2.2.4 Deep Mutual Learning (DML)

In the context of Deep Mutual Learning, as introduced by (Zhang et al. 2017), we utilized two student models, each based on the ResNet18 architecture. One model began its training with weights pre-initialized using the CIFAR-10/CIFAR-100 datasets, thus demonstrating enhanced performance in the initial phases of training due to its dataset-specific pretraining. In contrast, the second model initiated training with weights derived from the ImageNet dataset. This strategic variation in initial conditions was designed to capture diverse output probabilities and facilitate the learning of more nuanced knowledge, with the CIFAR-pretrained model also aimed at reducing training time by starting from a more advanced point of familiarity with the dataset.

As training progressed, we calculated and minimized the loss for both student models Θ_1 and Θ_2 using cross-entropy L_C and KL-divergence D_{KL}, the latter accounting for the soft targets provided by the peer student model, as shown in the following formulas:

$$L_{\Theta_1} = L_{C_1} + D_{KL}(p_2 \parallel p_1)$$
$$L_{\Theta_2} = L_{C_2} + D_{KL}(p_1 \parallel p_2)$$

These loss functions form the backbone of the Deep Mutual Learning strategy, where both models not only learn from the ground truth but also dynamically adjust their learning process based on the insights gained from each other's predictions.

Initially, we prioritized minimizing KL-divergence to quickly align the models' performance and accelerate the assimilation of diverse insights from the differing initial outputs. Later, the focus was shifted to cross-entropy loss to further enhance both models' performance. Following this mutual learning phase, the model that exhibited the highest test accuracy was selected for subsequent quantization experiments across various bit sizes.

2.2.5 Decoupled Knowledge Distillation (DKD)

We implemented the Decoupled Knowledge Distillation (DKD) method, introduced by Zhao et al. in CVPR 2022 (Zhao et al. 2022). DKD aims to address the imbalance between the logit-based distillation losses commonly used in knowledge distillation. Traditional methods often combine Kullback-Leibler (KL) divergence between the student and teacher logits with the standard cross-entropy loss, leading to suboptimal learning when the student's training objective becomes overly dependent on the teacher's output distribution. To mitigate this issue, DKD explicitly decouples the distillation loss into two components: target class knowledge distillation (TCKD) and non-target class knowledge distillation (NCKD). TCKD focuses on aligning the student's predicted probability of the correct class with the teacher's, ensuring the student learns the correct classification decision. NCKD, on the other hand, encourages the student to replicate the relative probability distribution of incorrect classes as predicted by the teacher. By adjusting the balance between these two components, DKD provides a more flexible and effective distillation process. This decoupling allows DKD to outperform conventional distillation methods, particularly in cases where the teacher provides overly confident predictions that may misguide the student. The method introduces two hyperparameters, alpha and beta, to control the contributions of TCKD and NCKD, enabling more fine-tuned knowledge transfer through adjusting the weight of these two components.

By experimenting with various baseline values for the α and β parameters, we were able to replicate some of the experiments from the original DKD paper. However, our results differed from the findings reported in the paper. Specifically, we tested α values of 0.5, 1, and 2, alongside β values of 4, 8, and 10. Our observations indicated a positive correlation between higher $\alpha - \beta$ combinations and improved test accuracy. The lowest accuracy, 86.68%, was observed with $\alpha = 0.5$ and $\beta = 10$, whereas the highest accuracy, 90.64%, was achieved with $\alpha = 2$ and $\beta = 10$ as shown in Figure 2. According to the paper, the best-performing configurations involved α values close to 1 and β values near 8. However, we were unable to replicate these results in our experiments. This discrepancy may stem from differences in our training setup. Notably, we utilized a pretrained ResNet-50 model sourced from Hugging Face, which may have introduced variations in initialization and learning dynamics. Additionally, differences in other hyperparameters and training conditions could have contributed to the observed deviations from the paper's reported performance.

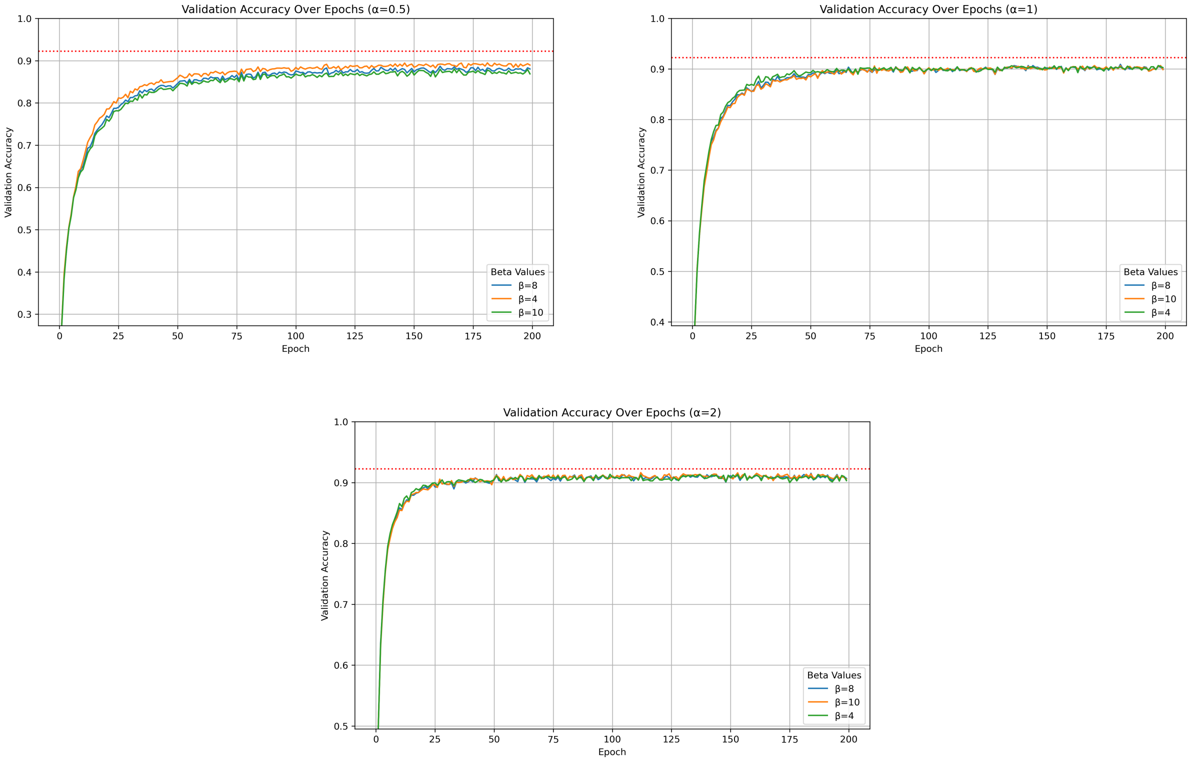

Figure 1: Validation accuracy for knowledge distillation across β values at each α

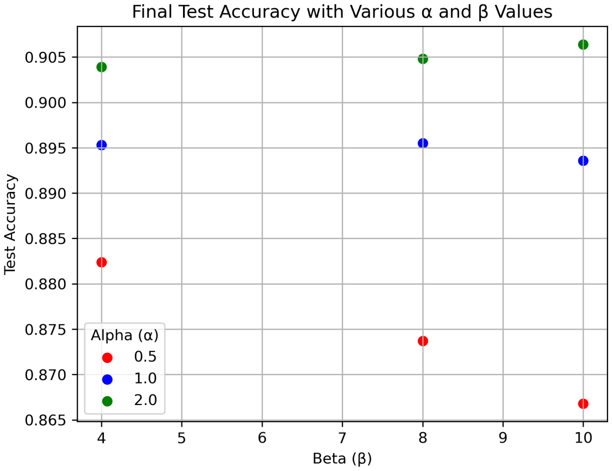

Figure 2: Final test accuracy of each model trained with different combination of α and β.

2.3 Post-Training Quantization

After distilling knowledge into the ResNet18 student model, we apply post-training quantization to further compress its storage footprint and accelerate inference. Our quantization strategy employs a greedy path-following algorithm to iteratively quantize layers to fewer bits. Once all layers are quantized, we evaluate the fully compressed model against the original student as well as the teacher model to compare accuracy, storage reduction systematically.

2.4 Evaluation Metrics

To thoroughly assess the effectiveness of our proposed approach, we employ two primary evaluation metrics:

Accuracy: We measure classification accuracy on benchmark datasets CIFAR-10 and CIFAR-100. Specifically, we focus on Top-1 accuracy, as it provides a direct and meaningful indicator of model performance for tasks with a relatively small number of classes, such as CIFAR-10. While metrics like Top-5 accuracy can offer additional insights for datasets with a larger number of classes, they become trivial in scenarios involving fewer classes. Hence, Top-1 accuracy is sufficient and more relevant for our analysis.

Compression Efficiency: We quantify the efficiency of the compressed models by evaluating the bit-width used to represent weights and activations post-quantization. Lower bit-width directly corresponds to smaller storage footprints and reduced computational requirements during inference. By systematically experimenting with different bit-widths—particularly 8-bit, 4-bit, and 2-bit quantization—we identify the optimal balance between storage efficiency and model accuracy. Moreover, we incorporate the total number of bits of the model to assess its efficiency, as we are not only comparing the distilled student model but also evaluating the teacher model and the student model together, which have different numbers of parameters. This provides a more comprehensive measure of compression efficiency.

Together, these metrics enable us to conduct a comprehensive comparative analysis, elucidating the trade-offs between model accuracy and compression efficiency.

3 Results

3.1 Experiment Results on CIFAR-10

3.1.1 Quantization Bit Width with Accuracy

From Table 1 and Figure 3, we observe that the teacher model maintains high accuracy even at very low bit-widths (2–3 bits), contrasting significantly with the student models, which initially experience pronounced accuracy drops at these lower bit widths. However, student models rapidly recover accuracy as bit widths increase, approaching their original accuracies at around 8 bits.

Among the four knowledge distillation methods evaluated, the Decoupled Knowledge Distillation (DKD) student experiences the steepest accuracy decline at lower bit-widths. Conversely, the Mixup method exhibits strong robustness starting at 3 bits, even surpassing the accuracy of the teacher model at higher bit-widths. Further analysis of this phenomenon is provided in Section 3.1.3.

Overall, the trends follow a diminishing returns pattern, where additional bit-width beyond 8 bits yields negligible accuracy improvements. Given the relatively simple nature of CIFAR-10, a significant portion of the information encoded at full precision (32 bits) is redundant, making lower-precision quantization models (8-bit) almost as effective as their full-precision counterparts.

Table 1: Accuracy for Various Models and Bit Sizes (CIFAR-10)

Model	2-bit	3-bit	4-bit	5-bit	6-bit	7-bit	8-bit	32-bit
Teacher	90.90	91.29	91.55	91.60	91.88	91.88	92.00	92.25
VKD Student	71.87	82.40	86.47	88.78	88.89	89.36	89.69	90.77
Mixup Student	88.12	92.63	94.2	95.09	95.21	95.18	95.45	95.77
DML Student	82.39	86.98	90.45	91.84	92.00	92.43	92.54	92.89
DKD Student	61.06	69.69	77.69	81.76	83.45	84.16	84.39	89.95

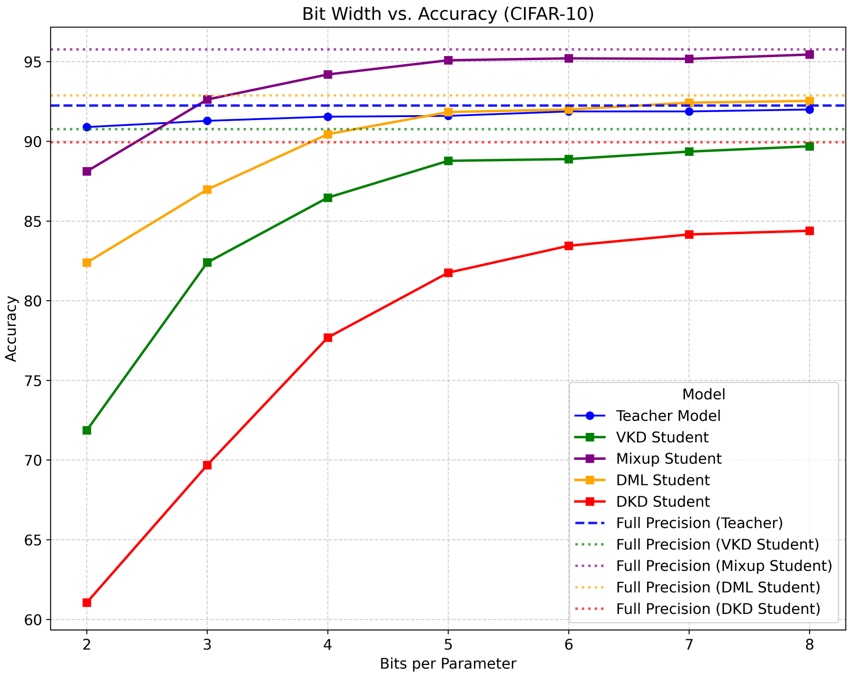

Figure 3: Accuracy vs. Bit Width on CIFAR-10.

3.1.2 Model Total Bits with Accuracy

Figure 4 illustrates that at comparable total model sizes (around 6 × 10^7 bits), student models distilled via Deep Mutual Learning (DML) and Mixup outperform the teacher model in terms of accuracy. This indicates that, particularly for simpler datasets like CIFAR-10, the integrated application of knowledge distillation and quantization significantly enhances model efficiency.

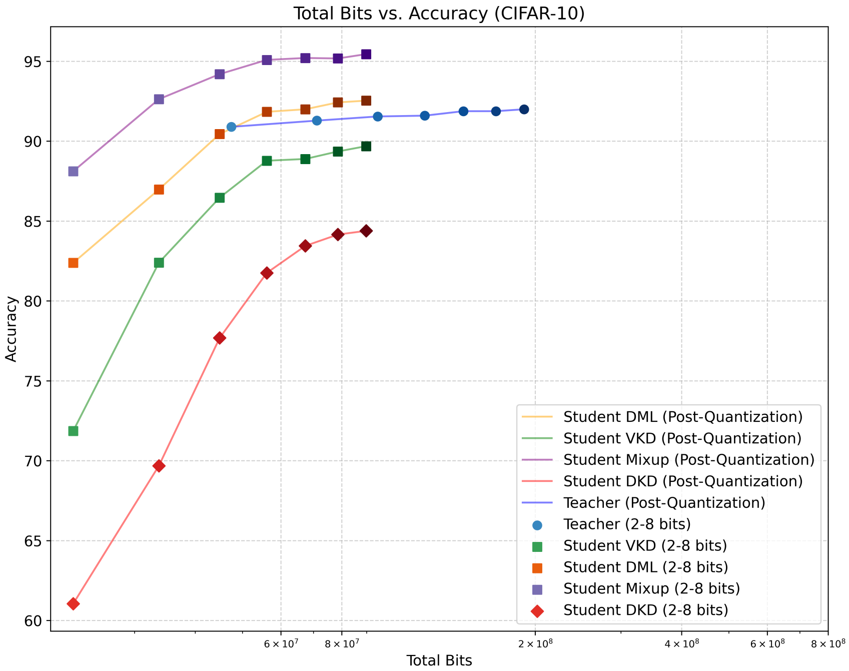

Figure 4: Accuracy vs. Model Total Bit on CIFAR-10

3.1.3 Justification on Student Model Outperforming Teacher Model

The notable observation of the Mixup student model surpassing the teacher model's accuracy on CIFAR-10 warrants further exploration. Mixup employs linear interpolation of image pairs and their labels, generating augmented data with enriched feature representations. This augmentation strategy likely contributes additional informative signals during the distillation process, enabling the student model to generalize better and achieve superior performance compared to the teacher model, especially given the simplicity and reduced overfitting risk inherent in CIFAR-10.

3.2 Experiment Results on CIFAR-100

3.2.1 Quantization Bit Width with Accuracy

As the bit width increases, the student models tend to increase rapidly, and the accuracy begins to converge at around 6 bits. Notably, the student models derived from Vanilla Knowledge Distilled (VKD) and Deep Mutual Learning (DML) exhibit distinct performance trends across different quantization levels. The VKD student model demonstrated a significant improvement from 26.23% at 2-bit to 68.70% at 6-bit, closely approaching the full-precision (32-bit) accuracy of 75.33%. This can also be observed in the DML student, which starts at a lower 19.22% at 2-bit but rapidly climbs to 70.87% at 8-bit.

In contrast, the Mixup student model shows a more modest performance increase, achieving 33.07% at 2-bit and plateauing around 62.16% from 7-bit onward, highlighting a potential limitation in its ability to leverage higher bit-width representations effectively. However, the student model derived from Decoupled Knowledge Distillation (DKD) lags behind the others at lower bit widths, with an accuracy of 22.20% at 2-bit and only reaching 52.85% at 8-bit. These results could be due to the significant number of classes present in the CIFAR-100 dataset, which also introduce more noise to the training of these models.

These results suggest that while lower-bit quantization severely degrades student model performance, increasing the bit width to at least 6-bit allows for significant recovery.

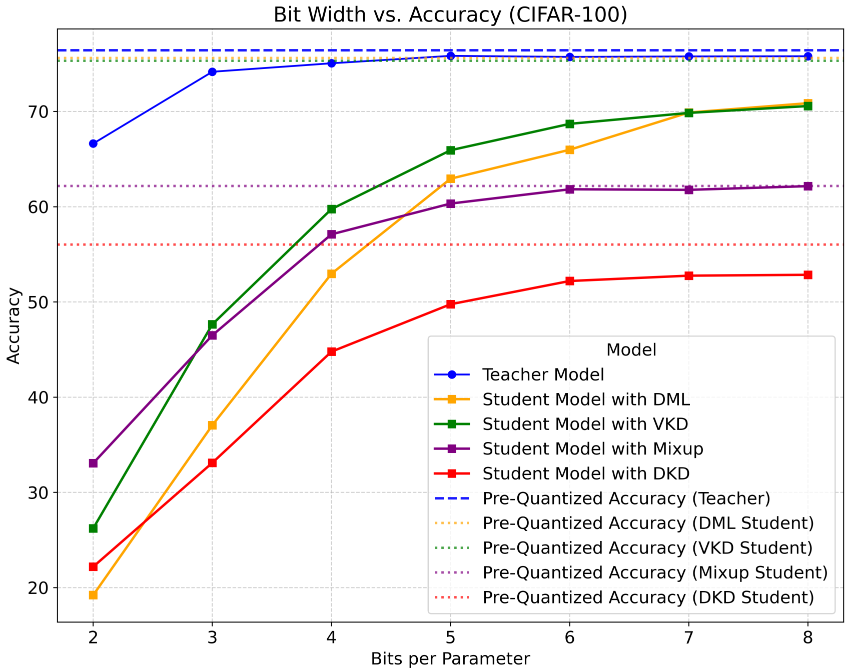

Figure 5: Accuracy vs. Bit Width on CIFAR-100

Table 2: Accuracy for Various Models and Bit Sizes (CIFAR-100)

Model	2-bit	3-bit	4-bit	5-bit	6-bit	7-bit	8-bit	32-bit
Teacher	66.63	74.17	75.06	75.84	75.73	75.78	75.80	76.43
VKD Student	26.23	47.65	59.75	65.93	68.70	69.85	70.54	75.33
Mixup Student	33.07	46.50	57.10	60.33	61.83	61.77	62.16	62.17
DML Student	19.22	37.06	52.96	62.94	65.98	69.89	70.87	75.12
DKD Student	22.20	33.11	44.78	49.76	52.20	52.76	52.85	58.01

3.2.2 Model Total Bits with Accuracy

From Figure 4, we observe that Vanilla Knowledge Distillation (VKD) and Deep Mutual Learning (DML) achieve the highest accuracies among the knowledge distillation methods and approach the accuracy of the teacher model at a significantly lower total bit count. However, at lower total bits, the performance of these models is significantly lower compared to the teacher at a similar total bit count. Similar to the quantization accuracy plot, the accuracy tends to converge, but in this case, it stabilizes closer to 16 bits, suggesting that the model benefits from a higher precision representation at this bit width.

Among the student models, VKD and DML exhibit the most rapid improvements in accuracy as the total bit count increases, reaching performance levels close to the teacher model. In contrast, the Mixup student model and the Decoupled Knowledge Distillation (DKD) student model show significantly lower accuracy across different bit widths and total bit allocations, consistent with the observations from Figure 3 and Table 2. The Mixup model converges at an accuracy of around 63%, which remains notably lower than the teacher model's performance. Similarly, the DKD model only achieves a final accuracy of approximately 56%, reinforcing the observation that it is less robust to quantization.

Interestingly, the final total bit count for all student models is significantly lower than that of the teacher model, demonstrating that despite their lower accuracy, the student models require fewer computational resources. More importantly, VKD and DML achieve comparable accuracy to the teacher while using significantly fewer total bits, highlighting their efficiency in terms of model compression and quantization. This suggests that while lower-bit quantization severely impacts student models at extreme bit reductions, increasing the total bit allocation to at least 16 bits enables a significant recovery in accuracy.

4 Discussion

The result section summarizes our findings on the CIFAR-10 and CIFAR-100 datasets, highlighting key differences in how knowledge distillation (KD) and quantization interact across varying levels of model complexity.

for CIFAR-10, our results indicate that the ResNet18 student model, distilled from ResNet50, maintains high accuracy across most quantization levels when trained on CIFAR-10. At

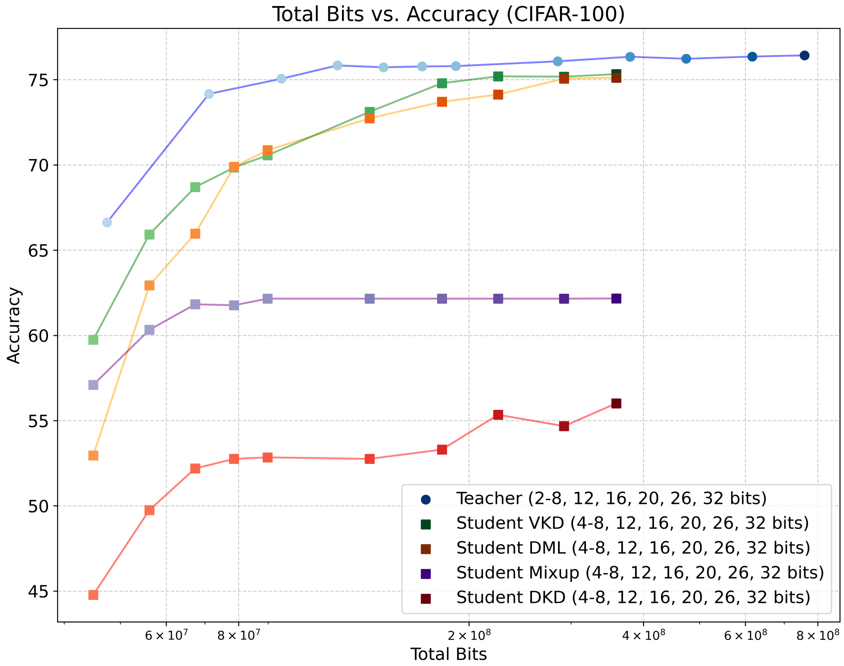

Figure 6: Accuracy vs. Model Total Bit on CIFAR-100

higher bit-widths (8-bit), the distilled model closely approximates the pre-quantization accuracy, reflecting effective knowledge transfer from the teacher model. Interestingly, accuracy only marginally declines as the quantization becomes more aggressive (e.g., 4-bit), signifying the robustness of our combined distillation and quantization method on simpler datasets.

The ResNet50 teacher model demonstrates notable stability, maintaining accuracy consistently across quantization levels, albeit slightly below its original baseline accuracy. This suggests that large architectures inherently possess higher resilience to precision reduction, whereas smaller distilled models benefit significantly from sequentially applied distillation and quantization.

However, when quantizing the student model to extremely low bit-widths (e.g., 2-bit), we observe a pronounced degradation in performance. This result indicates a critical threshold for bit-width reduction on models trained via knowledge distillation. Thus, practitioners aiming for deployment in severely resource-constrained environments should carefully consider this trade-off between accuracy and storage efficiency.

For CIFAR-100, our experiments highlight a more complex interplay between knowledge distillation and quantization. While the student model effectively captures general patterns through knowledge distillation, the complexity and richness of CIFAR-100 amplify accuracy sensitivity to bit-width reductions. Specifically, the accuracy drop is significantly steeper at quantization levels below 8 bits compared to CIFAR-10. The distilled model's accuracy remains relatively stable and competitive at moderate quantization (8-bit to 6-bit), but declines rapidly at lower bit-widths (4-bit and 2-bit), suggesting limitations in capturing finer-grained distinctions in more complex classification tasks.

Overall, our comparative analysis reveals that the integration of knowledge distillation and quantization methods significantly optimizes model compression for simpler datasets like CIFAR-10, achieving high accuracy even under aggressive quantization. For more challenging datasets like CIFAR-100, practitioners must adopt a cautious approach, balancing bit-width reductions against tolerable accuracy losses.

Future work includes investigating adaptive quantization strategies that dynamically adjust bit-widths at a layer-specific level and exploring hybrid methods integrating further compression techniques such as pruning to enhance performance at ultra-low bit-widths.

5 Conclusion

This study explored the impact of integrating quantization and knowledge distillation across two benchmark datasets, CIFAR-10 and CIFAR-100. For CIFAR-10, our findings demonstrate that distilled student models retain robustness across higher quantization levels (4-bit and above), closely matching pre-quantization performance. In contrast, for the more complex CIFAR-100 dataset, the distilled student models experience more significant accuracy degradation at lower quantization levels (2–6 bits). Consequently, while knowledge distillation provides considerable advantages in maintaining accuracy, it also imposes limitations on further compressibility through aggressive quantization, particularly in complex datasets. Overall, quantized teacher models tend to outperform quantized student models at extreme low-bit scenarios (e.g., 2 bits), though this advantage diminishes as dataset complexity decreases. Future research directions include exploring adaptive and layer-specific quantization strategies, and integrating additional compression methods such as pruning to further enhance model efficiency and accuracy at extremely low bit-widths.

References

Beyer, Lucas, Xiaohua Zhai, Amélie Royer, Larisa Markeeva, Rohan Anil, and Alexander Kolesnikov. 2022. “Knowledge distillation: A good teacher is patient and consistent.” [Link]

Bucilă, Cristian, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. “Model compression.” In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA Association for Computing Machinery. [Link]

Elthakeb, Ahmed T., Prannoy Pilligundla, Alex Cloninger, and Hadi Esmaeilzadeh. 2020. “Divide and Conquer: Leveraging Intermediate Feature Representations for Quantized Training of Neural Networks.” [Link]

Hinton, Geoffrey E., Oriol Vinyals, and Jeffrey Dean. 2015. “Distilling the Knowledge in a Neural Network.” *CoRR* abs/1503.02531. [Link]

Zhang, Jinjie, Yixuan Zhou, and Rayan Saab. 2023. “Post-training Quantization for Neural Networks with Provable Guarantees.” [Link]

Zhang, Ying, Tao Xiang, Timothy M. Hospedales, and Huchuan Lu. 2017. “Deep Mutual Learning.” [Link]

Zhao, Borui, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. 2022. “Decoupled Knowledge Distillation.” [Link]