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
Mentor: Alex Cloninger acloninger@ucsd.edu
Mentor: Rayan Saab rsaab@ucsd.edu

Halicioglu Data Science Institute, University of California - San Diego

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

Why model compression?

Deep Neural Networks (DNNs) have demonstrated outstanding performance in tasks ranging from image recognition to natural language processing.

- **Key Limitations:**
  - High computational demand that can hinder real-time inference and scalability.
  - Large storage requirements making deployment on resource-constrained devices challenging.
- **Common Compression Approaches:**
  - **Knowledge Distillation (KD):** Transfers learned representations from a large "teacher" model to a smaller "student" model.
  - **Quantization:** Lowers numerical precision of weights to reduce memory and computation costs.
- **Proposed Hybrid Solution:**
  - Integrates knowledge distillation and post-training quantization.
  - Optimizes the balance between model efficiency and predictive performance.

Research Question

Is quantizing a student model more efficient than quantizing the teacher model?

Data

- **CIFAR 10:** 60,000 images, 32x32 pixels, 10 classes.
- **CIFAR100:** 60,000 images, 32x32 pixels, 100 classes.

Methods

Track 1: Apply Knowledge Distillation to ResNet50, derive ResNet18, and quantize with Greedy Path-Following Quantization (GPFQ).

Track 2: Quantize ResNet50 using GPFQ.

- Knowledge Distillation Methods:
  - **Vanilla Knowledge Distillation**: Soft label + KL-divergence and cross entropy loss
  - **"Mixup" Method for Data Generation**: Interpolated training samples + KD
  - **Deep Mutual Learning**: Co-trained dual ResNet18 students + KD
  - **Decoupled Knowledge Distillation**: target-class (TCKD) loss + non-target-class (NCKD) loss
- Post-Training Quantization:
  - **Greedy Path-Following Quantization**: greedy layer-wise quantization to reduce bit size

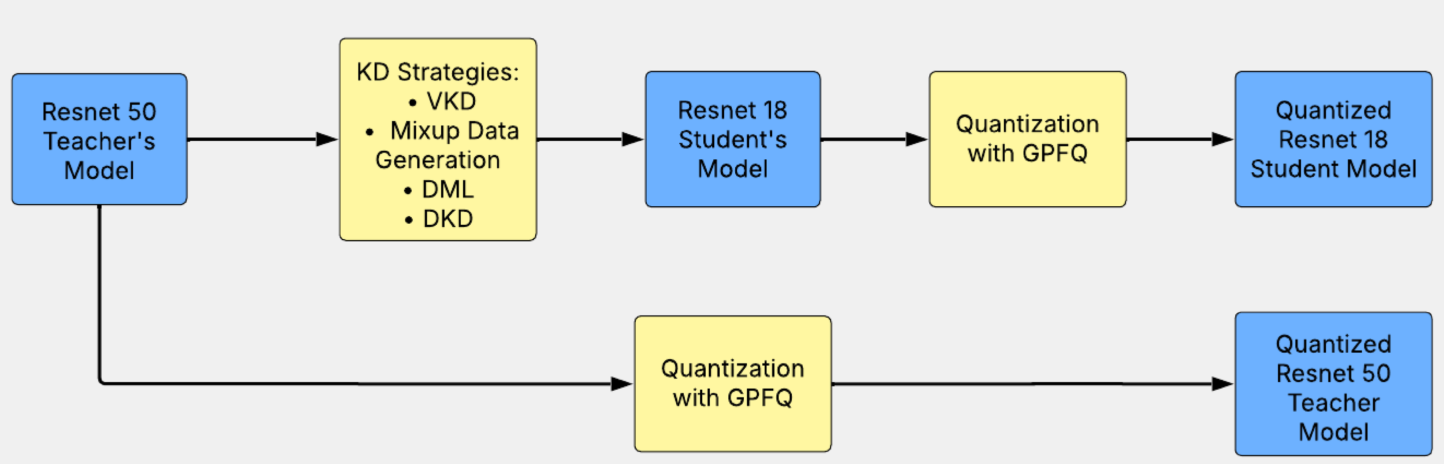

Figure 1. Flowchart for Experiment Design

Results

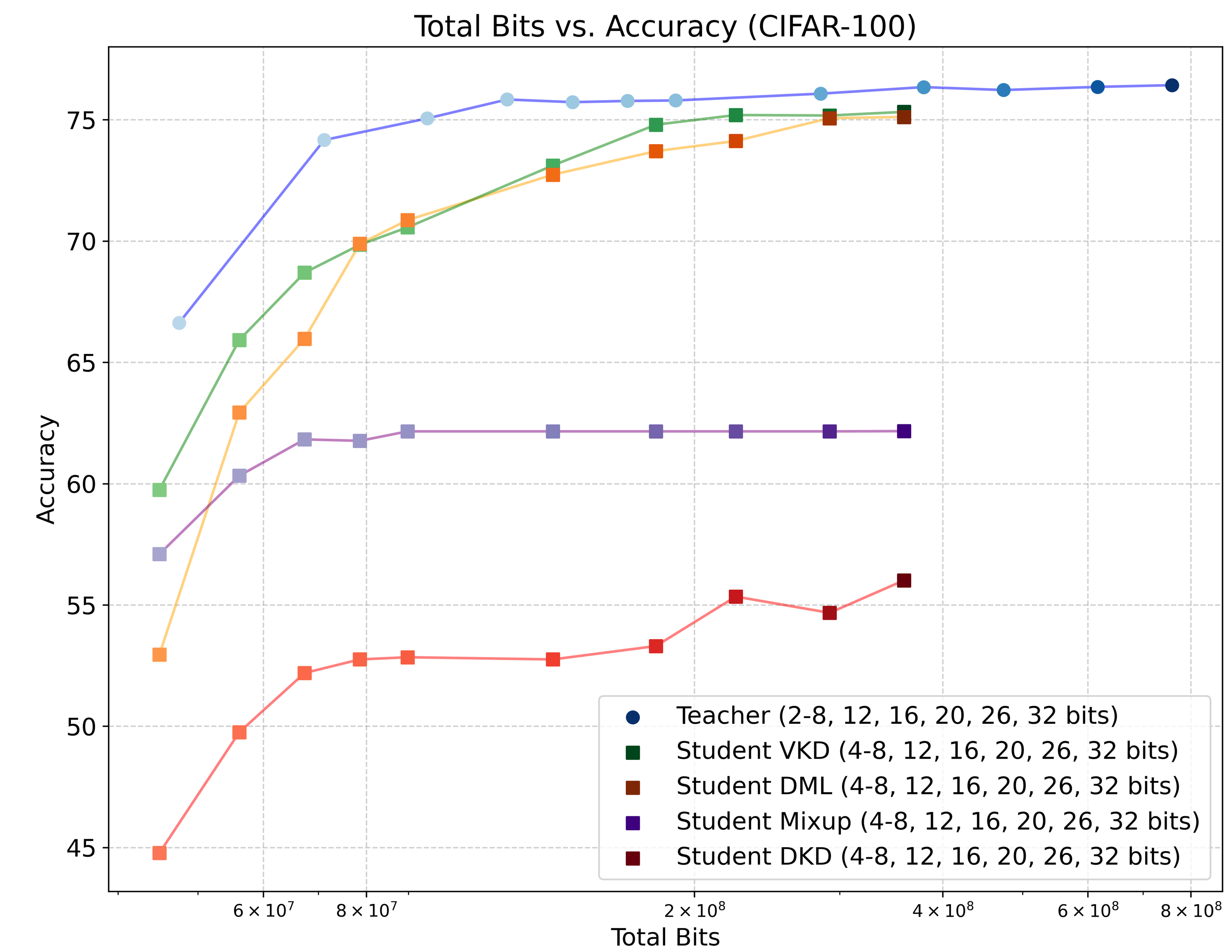

Figure 2. Accuracy vs. Memory

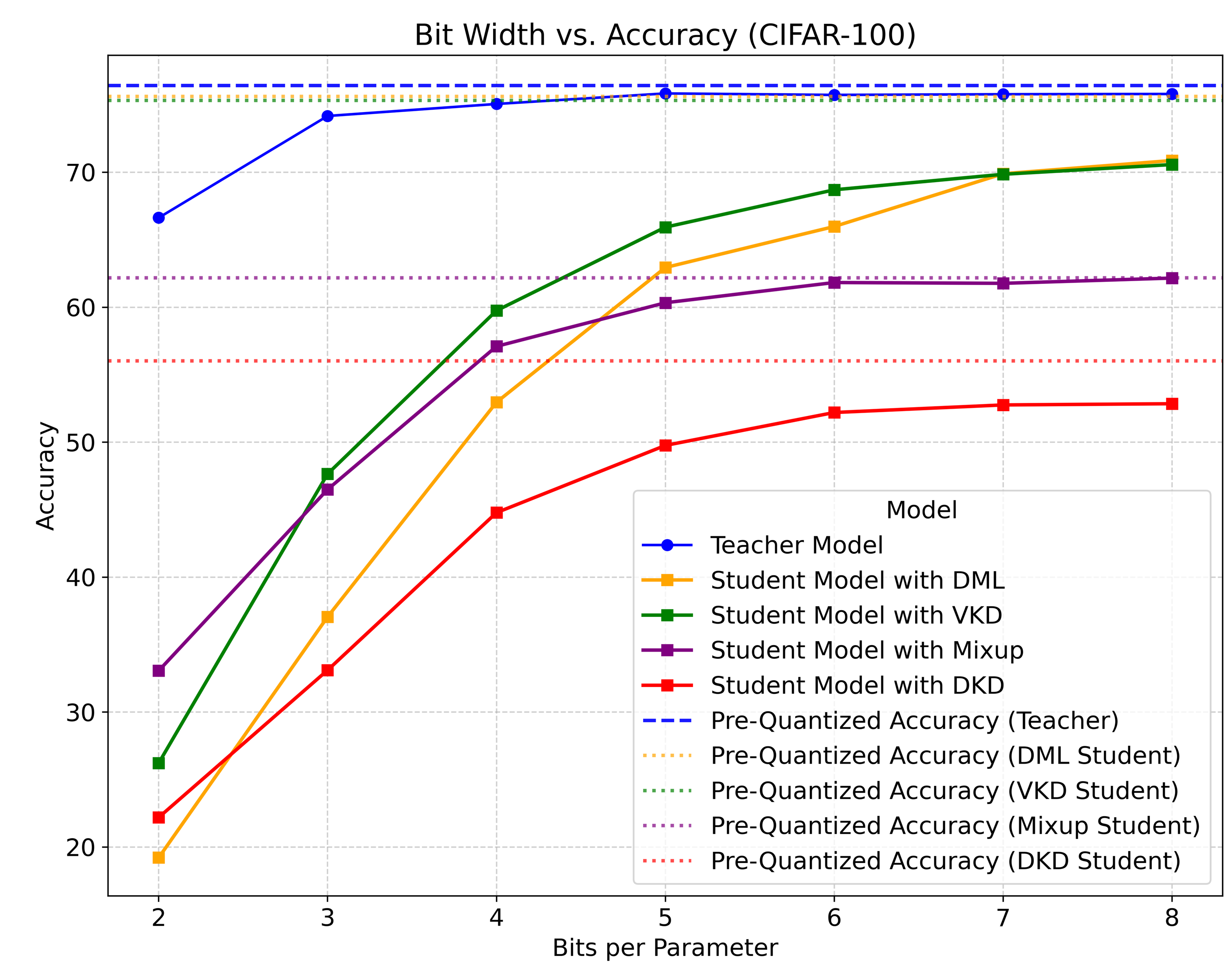

Figure 3. Accuracy vs. Quantization Bit Width

Results Contd.

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

Table 1. 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	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

Table 2. Accuracy for Various Models and Bit Sizes (CIFAR-10)

Conclusion

CIFAR-100:

- At 2–6 bits, all students initially drop in accuracy more sharply than the teacher.
- Distillation aids complex datasets but reduces compressibility for further quantization.

CIFAR-10:

- At 2–4 bits, all students see a sharper accuracy drop.
- Both models are robust at higher bit widths, with distillation helping the student nearly match its pre-quantization accuracy.

Overall, the quantized teacher model outperforms the quantized student model, particularly at small bit sizes, such as 2 bits. However, this advantage diminishes when applied to less complex datasets.

References

[1] C. B. et al., "Model compression," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '06. New York, NY, USA: Association for Computing Machinery, 2006, p. 535–541. [Online]. Available: https://doi.org/10.1145/1150402.1150464

[2] G. H. et al., "Distilling the knowledge in a neural network," *CoRR*, vol. abs/1503.02531, 2015. [Online]. Available: http://arxiv.org/abs/1503.02531

[3] A. R. et al., "Fitnets: Hints for thin deep nets," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. [Online]. Available: http://arxiv.org/abs/1412.6550

[4] J. Z. et al., "Post-training quantization for neural networks with provable guarantees," 2023. [Online]. Available: https://arxiv.org/abs/2201.11113

[5] Y. Z. et al., "Deep mutual learning," 2017. [Online]. Available: https://arxiv.org/abs/1706.00384

[6] L. B. et al., "Knowledge distillation: A good teacher is patient and consistent," 2022. [Online]. Available: https://arxiv.org/abs/2106.05237

[7] B. Z. et al., "Decoupled knowledge distillation," 2022. [Online]. Available: https://arxiv.org/abs/2203.08679