Improving Image Classification with Knowledge Distillation: A Student-Teacher Approach

Elliot Bu
University of Auckland
Auckland, New Zealand

***Abstract*—This report investigates the effectiveness of knowledge distillation (KD) as a model compression technique to enhance the efficiency of image classification tasks. A high-capacity ResNet34 model (teacher) is used to guide a more compact ResNet18 model (student), allowing the student to learn from both the ground truth labels and the soft target distributions generated by the teacher. The training process involves a weighted combination in Kullback-Leibler Divergence and cross-entropy loss, modulated by temperature scaling. Experimental results on the CIFAR-100 dataset demonstrate that the distilled student model achieves significantly improved accuracy compared to a baseline student model, while maintaining lower computational and memory requirements than the teacher. This balance of accuracy and efficiency makes the approach well-suited for real-world deployment in resource-constrained environments such as mobile and edge devices.**

Section 1. Introduction

Modern deep learning models excel in visual recognition tasks but often require substantial computational resources, making them impractical for real-time or edge-based applications. Knowledge distillation (KD), introduced by Hinton et al., addresses this issue by enabling a compact student model to learn from a larger, pre-trained teacher model. This report explores KD in the context of image classification, focusing on the 'ResNet' architecture and CIFAR-100 dataset. The primary objective is to evaluate whether KD can produce a lightweight model that retains strong performance. The report proceeds as follows: Section 2 covers related work, Section 3 outlines the methodology, Section 4 presents results, followed by discussion and conclusion in Sections 5 and 6.

Section 2. Literature Review

Knowledge distillation (KD), originally proposed by Hinton et al. [1], allows a small student model to learn from a larger teacher model by mimicking softened outputs. Romero et al. [2] extended this with FitNets, incorporating intermediate features. Zagoruyko and Komodakis [3] proposed attention transfer, aligning activation maps. Tung and Mori [4] introduced similarity-preserving KD to retain pairwise sample relationships. Urban et al. [5] explored Bayesian KD with uncertainty modelling. Mishra and Marr [6] developed Apprentice, enabling low-precision student training. While KD improves accuracy and efficiency, limitations include sensitivity to hyperparameters and limited analysis on mid-scale datasets like CIFAR-100. This project builds upon prior work by benchmarking KD with ResNet18 and ResNet34 on CIFAR-100.

Section 3. Methodology

In this report, I adopt a knowledge distillation framework comprising a ResNet34 model as the teacher and a smaller ResNet18 model as the student. The goal is to transfer the generalization capabilities of the teacher to the student, enabling efficient inference while preserving accuracy.

Section 3.1. Dataset and Preprocessing

I use the CIFAR-100 dataset, categorized into 10 classes (780 training images and 2600 test images, and 260 validation images per class). Data preprocessing includes standard normalization and data augmentation techniques, such as random cropping with padding and horizontal flipping, to improve model robustness and prevent overfitting.

Section 3.2. Training the Teacher Model

The teacher model (ResNet34) is trained from scratch using standard cross-entropy loss. The training process spans 10 epochs, using a learning rate of 0.001, batch size of 64, and SGD optimizer with a learning rate scheduler. Both training and validation losses and accuracies are recorded for performance tracking.

Section 3.3. Training the Student Model with Knowledge Distillation

The student model (ResNet18) is trained using a hybrid loss function that combines: Cross-entropy loss: for ground truth supervision. Kullback-Leibler (KL) divergence: to match the soft probability distributions output by the teacher. The final loss is defined as:

$$L_{total} = \alpha \cdot T^2 \cdot KL(\text{softmax}(\frac{Z_s}{T}), \text{softmax}(\frac{Z_t}{T})) + (1-\alpha) \cdot CE(Z_s, y)$$

Note:

- Z_s:Student model logits
- Z_t:Teacher model logits
- T=5: Temperature to soften predictions
- α=0.7: Weighting factor for KD loss vs. cross-entropy
- y: the ground truth label that the student should ultimately learn to predict accurately.

Hyperparameters: learning rate 0.001, batch size 64, 10 epochs. PyTorch was used for implementation and training was conducted on a GPU. Learning rate scheduling and validation checkpoints were applied.

Section 4. Results

The performance of both the teacher and student models was evaluated using training logs and validation metrics. The student model was trained under two configurations: with and without knowledge distillation (KD).

Section 4.1. Performance Curves

Fig. 1 shows the training and validation loss and accuracy curves for the student model (ResNet18) trained with knowledge distillation. The model demonstrates steady convergence and achieves higher validation accuracy compared to the baseline student.

Fig. 2 displays the performance of the teacher model (ResNet34), which exhibits strong generalization, although some overfitting is observed in later epochs.

Fig. 1: Loss and accuracy curves for ResNet18 (Student)

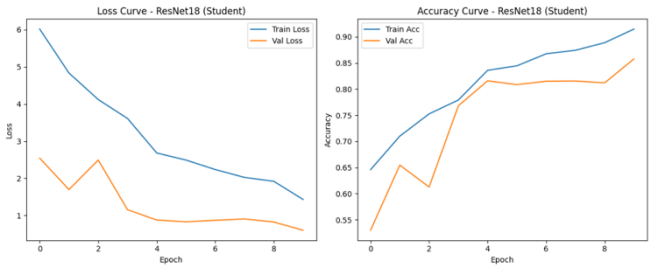
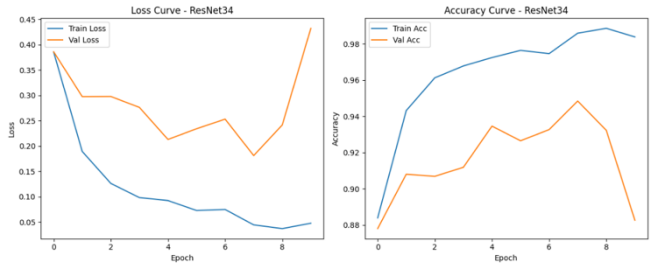


Fig. 2: Loss and accuracy curves for ResNet34 (Teacher)



Seciton 4.2 Accuracy Comparison

The best validation accuracy scores from all experimental runs are summarized in Table I. The student model trained using knowledge distillation (KD) achieved a validation accuracy of 85.76%, demonstrating strong performance while maintaining a significantly lower parameter count. In comparison, the teacher model (ResNet34) reached a higher accuracy of 94.85% but at nearly double the parameter cost. This trade-off highlights KD's effectiveness in compressing large models for efficient deployment without substantial accuracy degradation.

TABLE I. VALIDATION ACCURACY AND MODEL PARAMETERS COMPARISON

Model	*Val Accuracy*	*Params*	*Notes*
ResNet34	94.85%	21.8M	Teacher Model
ResNet18 (KD)	85.76%	11.2M	Student Model

The results confirm that knowledge distillation improves the student model's performance by approximately 9.09%, with only half the number of parameters compared to the teacher.

Section 5. Discussion

The results validate the primary objective of this project: to assess whether knowledge distillation (KD) can produce a compact yet high-performing student model for image classification. The KD-trained ResNet18 achieved a validation accuracy of 85.76%, demonstrating substantial performance while maintaining a significantly smaller model size compared to the ResNet34 teacher, which achieved 94.85%. This highlights KD's strength in transferring generalization ability from a large model to a smaller one, making it a practical solution for resource-constrained environments such as edge devices or mobile applications.

The training and validation curves further support the effectiveness of KD. The student model exhibited steady learning and generalization without signs of severe overfitting, unlike the teacher model, which showed some overfitting in later epochs. These findings suggest that KD not only improves accuracy but may also introduce beneficial regularization.

However, limitations remain. The results are based solely on the CIFAR-100 dataset, which, while diverse, is relatively small and balanced. The generalizability of these findings to larger or more complex datasets—such as those with class imbalance or higher resolution images—has yet to be explored. The performance of KD is sensitive to hyperparameters like temperature (T) and distillation weight (α), requiring careful tuning to achieve optimal results.

Through this project, we gained a deeper understanding of how soft targets influence learning and how model compression techniques like KD can balance performance and efficiency. Future work could explore combining KD with other compression strategies, such as pruning or quantization, or applying the approach to more complex real-world tasks.

SECTION 6. Conclusion

This project demonstrates the effectiveness of knowledge distillation (KD) as a model compression strategy for image classification. By transferring knowledge from a high-capacity ResNet34 teacher to a lightweight ResNet18 student, we achieved strong performance—85.76% validation accuracy—with nearly half the parameters of the teacher model. This validates that KD can effectively bridge the performance gap while significantly reducing computational cost, making the approach highly suitable for deployment on edge devices or resource-limited environments.

The key contribution of this work lies in showing that soft targets from a well-trained teacher can guide a smaller network to learn richer representations than traditional supervised learning alone. This confirms the potential of KD to improve generalization without increasing model size.

For future research, exploring KD on more complex or imbalanced datasets, integrating it with pruning or quantization techniques, or extending it to self-distillation and multi-teacher strategies could further enhance both

performance and efficiency. Refining hyperparameter tuning for temperature and loss weighting may yield further gains.

This project highlights KD as a practical and scalable method for building efficient deep learning systems without compromising much on accuracy.

SECTION 7. REFERENCES

[1] Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the Knowledge in a Neural Network.

[2] Romero, A., Ballas, N., Kahou, S. E., et al. (2014). FitNets: Hints for Thin Deep Nets.

[3] Zagoruyko, S., & Komodakis, N. (2017). Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer.

[4] Tung, F., & Mori, G. (2019). Similarity-Preserving Knowledge Distillation.

[5] Urban, G., et al. (2016). Do Deep Convolutional Nets Really Need to be Deep and Convolutional?

[6] Mishra, A., & Marr, D. (2018). Apprentice: Using Knowledge Distillation Techniques to Improve Low-Precision Network Accuracy.