SAPIENZA
UNIVERSITÀ DI ROMA

Department of Computer, Control and Management Engineering (DIAG)

Confidence Aware Ensemble Knowledge Distillation

Advanced Machine Learning
Final Project Report

Professor:
Fabio Galasso

Students:
Abzal Aidakhmetov
Adilkhan Bakridenov
Eldar Gabdulsattarov
Nadir Nuralin

Academic Year 2024/2025

Contents

1 Introduction

Knowledge distillation (KD) enables lightweight models to achieve high performance by learning from larger teacher models, making it essential for resource-constrained environments like mobile devices and IoT. The core idea is to transfer knowledge through the teacher's logits, providing rich data patterns beyond ground truth labels.

This project focuses on response-based KD, which uses only the last layer's logits [1], addressing challenges in multi-teacher scenarios such as balancing diverse teacher contributions and mitigating noisy predictions. Using MobileNet V2 as the student and a combination of ResNet-50, DenseNet-121, and ViT-B16 as teachers, we propose three novel methods: confidence-aware weighted KD, dynamic alpha-guided KD, and adaptive alpha-guided KD.

Experiments conducted on CIFAR-100 demonstrate the potential for improving distillation efficiency and student model accuracy under computational constraints.

2 Methodology

The CIFAR-100 dataset was selected for experimentation. This dataset consists of 60,000 32x32 color images divided into 100 classes. The data was split into 45,000 training images, 5,000 validation images, and 10,000 test images. The preprocessing steps ensured that the dataset was suitable for the classification task.

For the student model, we used MobileNet V2, which is a lightweight convolutional neural network with 2.4 million parameters [2], known for its efficiency in resource-constrained environments. Teacher models included ResNet-50, DenseNet-121, and ViT-B16/224. These models were chosen for their diverse architectures, representing convolutional neural networks and transformers, with parameter counts of 25.6 million, 7.98 million and 86 million, respectively.

Training used the AdamW optimizer with a learning rate of 1e-2. Cross-entropy and KL divergence loss functions were used to balance teacher predictions and ground truth labels. The implementation was based on PyTorch. According to the findings in [3], the student model benefits from long-term, consistent, patient training to match the performance of the teacher model. However, due to computational constraints, all our experiments are limited to 100 epochs and executed on NVIDIA A100 GPUs. Due to limited computational resources, we prioritized temperature scaling to smooth teacher logits and alpha weighting to balance the guidance provided by teachers and the dataset's ground truth.

2.1 Single Teacher KD

The loss function for KD with a single teacher is defined as:

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{KL} + (1 - \alpha) \cdot \mathcal{L}_{CE}$$

where:

$$\mathcal{L}_{KL} = \text{KLDiv}\left(\log\left(\sigma\left(z_S\right)\right), \sigma\left(z_T\right)\right) \cdot \tau^2$$

Here:

- KLDiv is the Kullback-Leibler (KL) divergence, which measures the difference between the softened logits of the student (z_S) and the teacher (z_T).
- $\mathcal{L}_{CE}$ is the cross-entropy loss, calculated between the student's predictions and the ground truth.
- σ represents the softmax function, applied with a temperature τ to control the smoothness of the output logits:

$$\sigma(z^c) = \frac{\exp(z^c/\tau)}{\sum_j \exp(z^j/\tau)}.$$

This formulation combines the knowledge transfer from the teacher via $\mathcal{L}_{KL}$ with the standard supervised learning from ground truth labels via $\mathcal{L}_{CE}$, with the trade-off controlled by the coefficient α (set to 0.5).

2.2 Confidence-Aware Multi-Teacher KD

The diagram in Fig. 1 shows a multi-teacher distillation process where the **KL divergence** (L_{KL}) aligns the student's logits with a weighted sum of teachers' logits, and the **cross-entropy** (L_{CE}) aligns the student's predictions with the ground truth. Combining these losses gives the total loss for training.

We define the labeled training dataset as $\{(x_i, y_i)\}_{i=1}^N$, where N is the total number of samples, and K is the number of teacher models. Let $\mathbf{z} = [z^1, \ldots, z^C]$ denote the logits, where C is the number of classes.

In the subsequent sections, we provide a detailed explanation of the Confidence-Aware Multi-Teacher KD methods.

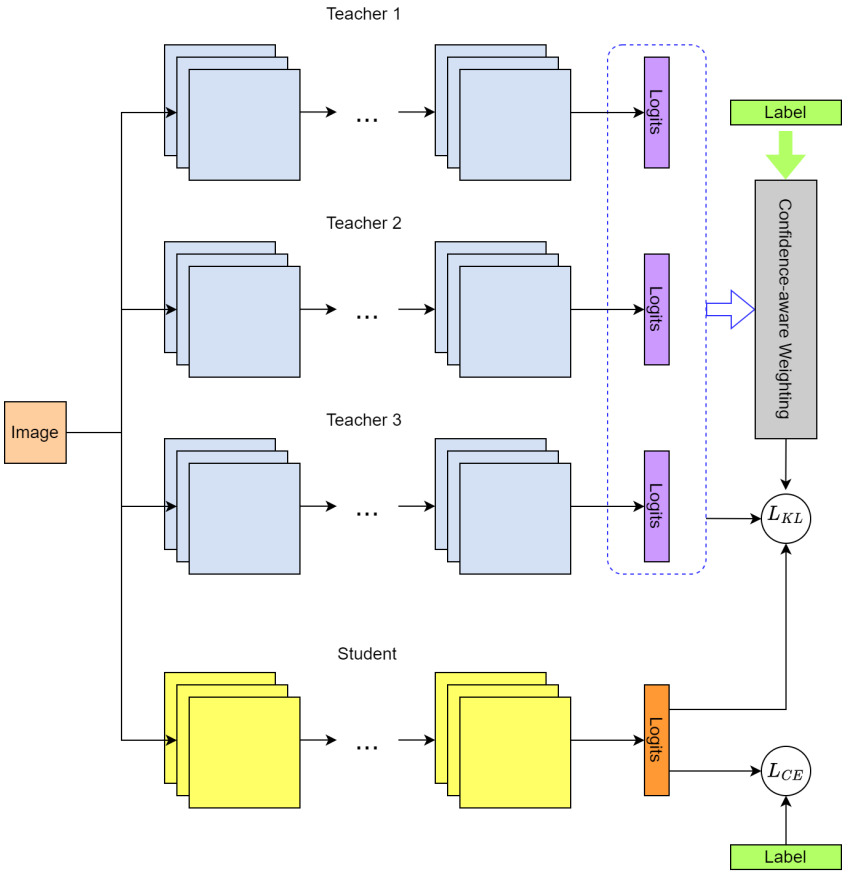

Figure 1: Illustration of the AML Multi-Teacher approach.

2.2.1 Method 1: Confidence-Aware Weighted KD (CA-WKD)

We have been inspired by [4] and implemented the weighting formula in Eq. 1. This method is designed to assign a higher weight to teachers with higher probabilities for the ground truth class.

$$\mathcal{L}^k_{CE_{KD}} = -\sum_{c=1}^{C} y^c \log\left(\sigma\left(z^c_{T_k}\right)\right)$$

$$w^k_{KD} = \frac{1}{K-1}\left(1 - \frac{\exp\left(\mathcal{L}^k_{CE_{KD}}\right)}{\sum_j \exp\left(\mathcal{L}^j_{CE_{KD}}\right)}\right) \tag{1}$$

$$\mathcal{L}^k_{KL} = \text{KLDiv}\left(\log\left(\sigma\left(z_S\right)\right), \sigma\left(z_{T_k}\right)\right) \cdot \tau^2 \tag{2}$$

$$\mathcal{L}_{KL} = \sum_{k=1}^{K} w^k \cdot \mathcal{L}^k_{KL} \tag{3}$$

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{KL} + \mathcal{L}_{CE} \tag{4}$$

The total loss ($\mathcal{L}_{\text{Total}}$) is a combination of the KL loss and the CE loss, both of which have the same influence on the student model. One drawback of this method is that when all teachers are incorrect, their weights will not be small enough, causing the student to learn from weak teachers.

2.2.2 Method 2: α-Guided CA-WKD

To address the issue in Method 1, we introduced the coefficient α, which dynamically adjusts the balance between the cross-entropy loss ($\mathcal{L}_{CE}$) and the Kullback–Leibler divergence loss ($\mathcal{L}_{KL}$).

- When teachers are weak (i.e., misclassifying), α gives more preference to $\mathcal{L}_{CE}$, relying more on the ground truth labels.

- When teachers are strong (i.e., classifying correctly), α prioritizes $\mathcal{L}_{KL}$, allowing the student to learn effectively from the teachers' predictions.

The total loss is computed as:

$$\mathcal{L}_{\text{Total}} = \alpha \cdot \mathcal{L}_{KL} + (1 - \alpha) \cdot \mathcal{L}_{CE} \tag{5}$$

$$\alpha = \sigma_{sigmoid}\left(\gamma \cdot (\text{Conf}_{\text{avg}} - \theta)\right) \tag{6}$$

where:

- $\sigma_{\text{sigmoid}}(x)$: Sigmoid activation function.
- γ: Scaling factor, with a default value of 2.
- Conf_{avg}: The average confidence, calculated by taking the mean logits of all teachers, passing them through a softmax, and averaging over the batch of images.
- θ: A threshold value with default value of 0.3.

This approach reduces the influence of teachers that misclassify the input, ensuring that the distillation process emphasizes correct predictions and improves the overall performance of the student model.

2.2.3 Method 3: Adaptive α-Guided KD

In this method, we address the limitations of Method 2 by dynamically computing α in a closed-form expression. Unlike Method 2, where α depends on fixed hyperparameters (γ and θ), this approach eliminates the need for hyperparameter tuning, which can be computationally expensive in KD tasks.

The coefficient α is defined as:

$$\alpha_i = \max_{k \in \{1,2,\ldots,K\}} p_{T_k}(y_i \mid x_i),$$

where $p_{T_k}(y_i \mid x_i)$ is the probability that teacher T_k assigns to the true class y_i for input x_i. This value of α is then averaged across the batch and applied to the standard formula for KD.

The total loss is computed as:

$$\mathcal{L}_{\text{Total}} = \alpha \cdot \mathcal{L}^*_{KL} + (1 - \alpha) \cdot \mathcal{L}_{\text{CE}},$$

where:

- $\mathcal{L}^*_{KL}$ represents the KL divergence loss calculated using the predictions from the most confident teacher.

Instead of taking a weighted sum of the logits from all teachers for $\mathcal{L}_{KL}$, we select the most confident teacher, i.e., the teacher that assigns the highest probability to the correct class. This ensures that the student primarily learns from the strongest predictions and avoids the influence of weaker teachers. By using the most confident teacher's prediction for α, we also avoid introducing additional hyperparameters, simplifying the training process while maintaining robust performance.

3 Results and Discussion

This project investigated multi-teacher KD to enhance the performance of a student model, MobileNet V2, by leveraging multiple larger teacher models. As illustrated in Fig. 2, the baseline student model was trained without KD, achieving the lowest validation, as well as test accuracy of 66% shown in Table 1. Subsequently, single-teacher KD was performed using ResNet-50, DenseNet-121, and ViT-B16/224, resulting in student accuracies between 73% and 75%, indicating significant improvements over the baseline.

We then implemented a multi-teacher KD approach, employing confidence-aware weighting mechanisms to mitigate noisy predictions by assigning distinct weights to each teacher. While this ensemble method slightly improved the student model's performance, the gains were less substantial compared to the best single-teacher results.

Evaluation focused on accuracy and loss reduction, considering computational efficiency. Notably, DenseNet-121, despite having a teacher accuracy of 76%, enabled the student to achieve the highest single-teacher distilled accuracy of 75.38%. In contrast, the multi-teacher ensemble provided marginal improvements but did not surpass DenseNet-121's standalone performance, highlighting the complexities and trade-offs in multi-teacher distillation.

3.1 Challenges and Limitations

A primary challenge was limited computational resources, restricting the number of training epochs and the extent of hyperparameter tuning. These constraints likely hindered the multi-teacher framework from realizing its full potential, particularly where extended training and finely tuned parameters could enhance performance [3].

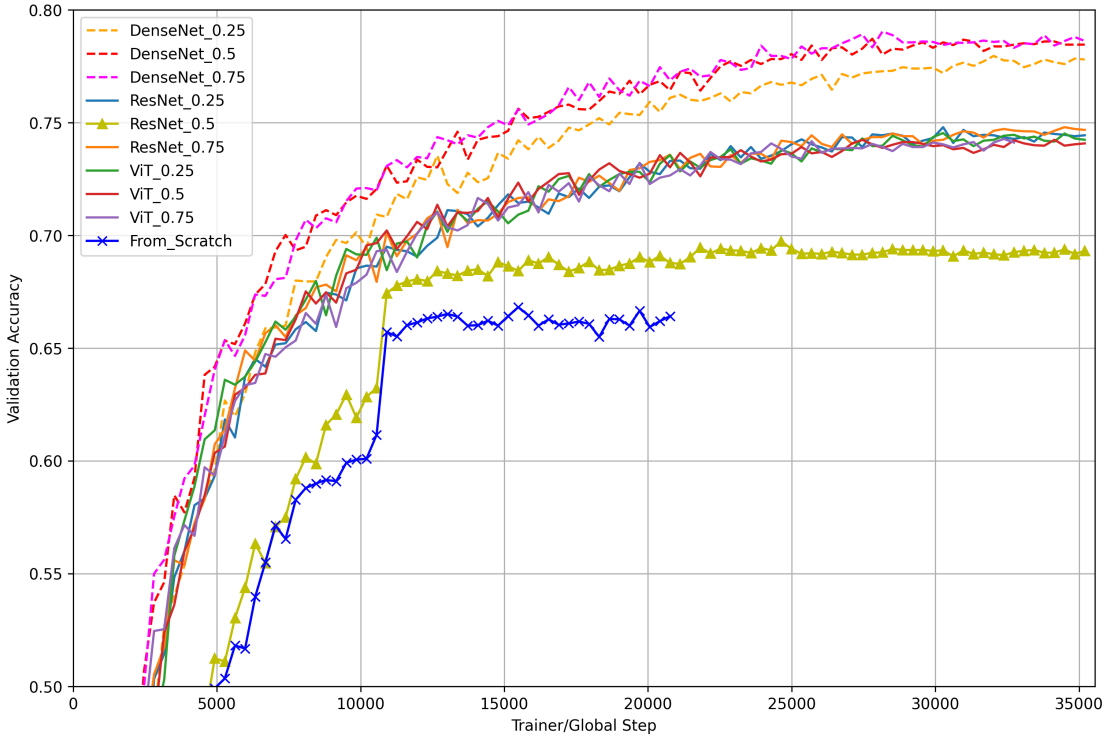

Figure 2: Comparison of Validation Accuracy for Single Teacher KD.

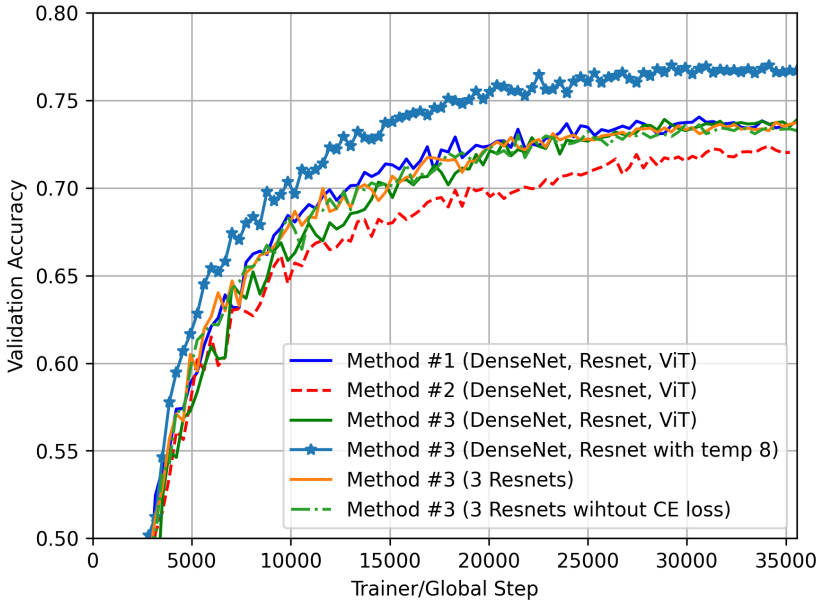

Figure 3: Comparison of Validation Accuracy for Multi-Teacher KD.

3.2 Performance Insights

For multi-teacher KD, three methods were explored, with validation accuracies depicted in Fig. 3:

1. **CA-WKD (Method 1)**: Assigning different weights to each teacher yielded a student accuracy of 73.33%. However, this method shared weights even when all teachers were incorrect.

2. **α-Guided CA-WKD (Method 2)**: Introducing a dynamic alpha to adjust each teacher's influence based on performance did not improve accuracy.

3. **Adaptive α-Guided KD (Method 3)**: Utilizing the most confident teacher's prediction as the alpha value enhanced the multi-teacher network's accuracy to 74%. Fig. 4 illustrates that higher temperatures during training improved validation and test accuracies, likely due to distribution smoothing.

Additionally, using three different ResNet models (ResNet18, ResNet34 and ResNet50) as teachers revealed that with Method 3, the loss function was heavily influenced by the KL divergence due to teacher confidence. Omitting the CE

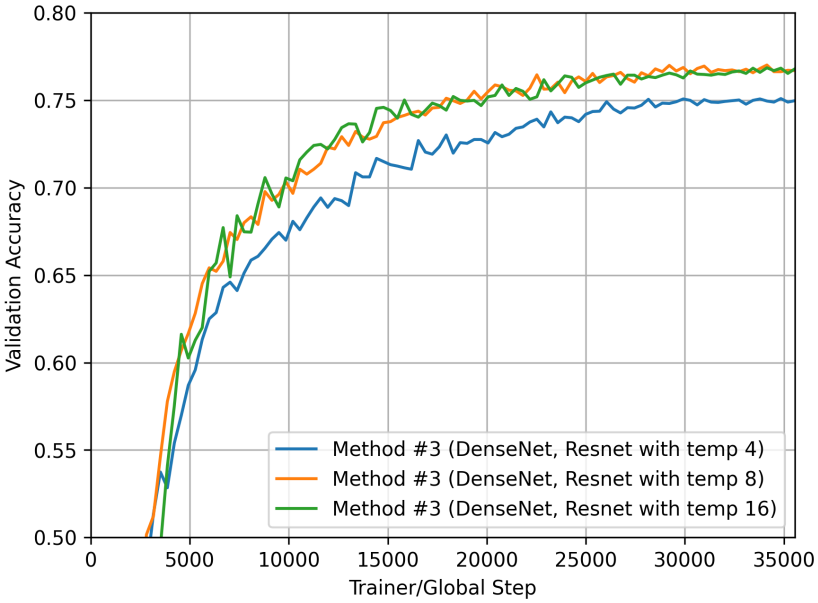

Figure 4: Temperature Hyperparameter Tuning (Multi-Teacher KD).

loss and training solely with KL loss slightly decreased accuracy from 72.90% to 71.89%, but improved convergence by simplifying the loss function.

Training Type	**Accuracy**
Baseline (Student from Scratch)	66.29%
Single Teacher (ResNet w/ $\alpha = 0.75$)	73.75%
Single Teacher (ViT w/ $\alpha = 0.25$)	74.00%
Single Teacher (DenseNet w/ $\alpha = 0.50$)	75.38%
Multi-Teacher Method 1 (DenseNet, ResNet & ViT)	73.33%
Multi-Teacher Method 2 (DenseNet, ResNet & ViT)	71.38%
Multi-Teacher Method 3 (DenseNet, ResNet & ViT)	72.10%
Multi-Teacher Method 3 (DenseNet, ResNet w/ temp 8.0)	74.00%
Multi-Teacher Method 3 (DenseNet, ResNet w/ temp 16.0)	73.96%
Multi-Teacher Method 3 (3 ResNets)	72.90%
Multi-Teacher Method 3 (3 ResNets without CE loss)	71.89%

Table 1: Test Accuracies of Different Training Methods

4 Conclusion

This study introduces a confidence-aware multi-teacher KD framework, leveraging novel weighting and alpha-guided methods to improve student model accuracy on CIFAR-100 while maintaining efficiency. Single-teacher KD with DenseNet and multi-teacher KD with temperature 8 achieved the highest results, both outperforming the baseline. The findings suggest that KD consistently improves performance, with hyperparameter tuning and extended training potentially enabling students to match teacher performance, especially for smaller teacher models like DenseNet. Additionally, the results indicate that soft loss has a greater impact on student performance than CE loss, with similar outcomes observed when CE is excluded.

References

[1] Jing Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129:1789–1819, 2021.

[2] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Inverted residuals and linear bottlenecks: Mobile networks forclassification, detection and segmentation. 01 2018.

[3] Lucas Beyer, Xiaohua Zhai, Amelie Royer, Larisa Markeeva, Rohan Anil, and Alexander Kolesnikov. Knowledge distillation: A good teacher is patient and consistent. pages 10915–10924, 06 2022.

[4] Hailin Zhang, Defang Chen, and Can Wang. Confidence-aware multi-teacher knowledge distillation. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4498–4502, 2022.