CS6302: Advanced Topics In Machine Learning

Deep Model Compression: Knowledge Distillation Techniques

Ahsan Mir
25100325@lums.edu.pk

Lahore University of Management Sciences

1 Knowledge Distillation

1.1 Mathematical Proof for Logit Matching

In this section, we use the basic logit matching equation for knowledge distillation to derive the following expression outlined in Hinton's Approach:

$$\frac{\partial L_{KL}}{\partial z_i} \approx \frac{1}{NT^2}(z_i - v_i)$$

For the purpose of this proof, the following definitions are used:

- z_i: i-th logit (pre-softmax output) of the student model.
- v_i: i-th logit of the teacher model.
- q_i: Soft probability produced by the student model for the i-th category (channel).
- p_i: Soft probability produced by the teacher model for the i-th category (channel).
- T: Temperature parameter used to soften the probabilities.
- N: Total number of logits (categories/classes).

I. **Distillation Loss Function L_{KL}**

The distillation loss when using logit matching is based on the Kullback-Leibler (KL) divergence between the teacher's and student's output probabilities:

$$L_{KL} = \sum_i p_i \log\left(\frac{p_i}{q_i}\right)$$

Since p_i is independent of the student's logits z_i (as p_i comes from the teacher model), when computing derivatives with respect to z_i, the term involving $p_i \log p_i$ can be treated as a constant and ignored. Therefore, the loss simplifies to:

$$L_{KL} = -\sum_i p_i \log q_i$$

II. **Soft Probabilities (q_i, p_i)**

The soft probabilities q_i and p_i are calculated using the softmax function:

$$q_i = \frac{e^{z_i/T}}{\sum_j e^{z_j/T}}, \quad p_i = \frac{e^{v_i/T}}{\sum_j e^{v_j/T}}$$

III. **Loss Gradient $\frac{\partial L_{KL}}{\partial z_i}$**

The gradient (partial derivative) of the loss L_{KL} with respect to the student's logits z_i is:

$$\frac{\partial L_{KL}}{\partial z_i} = -\sum_k p_k \frac{\partial \log q_k}{\partial z_i}$$

To find this gradient, we first need to compute $\frac{\partial \log q_k}{\partial z_i}$. Using the quotient rule for logarithms, we know that:

$$\log q_k = \frac{z_k}{T} - \log\left(\sum_j e^{z_j/T}\right)$$

Therefore, the derivative is:

$$\frac{\partial \log q_k}{\partial z_i} = \frac{1}{T}\delta_{ki} - \frac{1}{T}\frac{e^{z_i/T}}{\sum_j e^{z_j/T}} = \frac{1}{T}(\delta_{ki} - q_i)$$

Here, δ_{ki} is the Kronecker delta function, which equals 1 if $k = i$ and 0 otherwise.

Substituting back into the gradient:

$$\frac{\partial L_{KL}}{\partial z_i} = -\frac{1}{T}\sum_k p_k(\delta_{ki} - q_i) = -\frac{1}{T}(p_i - q_i)$$

Simplifying, we get:

$$\frac{\partial L_{KL}}{\partial z_i} = \frac{1}{T}(q_i - p_i)$$

IV. **Approximation of soft probabilities for large temperature T**

Assuming that T is large compared to the magnitude of the logits (z_i and v_i), we can use the first-order Taylor expansion for the exponential function. Recall that for small x:

$$e^x \approx 1 + x$$

Applying this to $e^{z_i/T}$ and $e^{v_i/T}$, we get:

$$e^{z_i/T} \approx 1 + \frac{z_i}{T}, \quad e^{v_i/T} \approx 1 + \frac{v_i}{T}$$

Next, we compute the denominators:

$$\sum_j e^{z_j/T} \approx \sum_j \left(1 + \frac{z_j}{T}\right) = N + \frac{1}{T}\sum_j z_j$$

Similarly for $\sum_j e^{v_j/T}$, we get:

$$\sum_j e^{v_j/T} \approx N + \frac{1}{T}\sum_j v_j$$

Given the assumption that the logits are zero-meaned separately for each transfer case:

$$\sum_j z_j = 0, \quad \sum_j v_j = 0$$

Thus, the denominators we have computed simplify to:

$$\sum_j e^{z_j/T} \approx N, \quad \sum_j e^{v_j/T} \approx N$$

Using these results, we can approximate q_i and p_i as follows:

$$q_i \approx \frac{1+\frac{z_i}{T}}{N}, \quad p_i \approx \frac{1+\frac{v_i}{T}}{N}$$

V. **Derivation**

Using the approximations from the previous step, we can compute the difference of our soft probabilities, $q_i - p_i$, to expand our gradient expression:

$$q_i - p_i \approx \frac{1+\frac{z_i}{T}}{N} - \frac{1+\frac{v_i}{T}}{N} = \frac{1}{N}\left(\frac{z_i}{T} - \frac{v_i}{T}\right) = \frac{1}{NT}(z_i - v_i)$$

Substituting back into the gradient and simplifying the expression, we achieve the desired result:

$$\frac{\partial L_{KL}}{\partial z_i} = \frac{1}{T}(q_i - p_i) \approx \frac{1}{T}\left(\frac{1}{NT}(z_i - v_i)\right) = \frac{1}{NT^2}(z_i - v_i)$$

VI. **Conclusion**

Under the assumptions:

- **High Temperature Approximation**: The temperature T is large compared to the magnitudes of the logits z_i and v_i, allowing us to use the first-order Taylor expansion of the exponential function.
- **Zero-Meaned Logits**: The logits have been zero-meaned separately for each transfer case, so $\sum_j z_j = 0$ and $\sum_j v_j = 0$.

We have derived:

$$\frac{\partial L_{KL}}{\partial z_i} \approx \frac{1}{NT^2}(z_i - v_i)$$

This result indicates that, in the high-temperature limit, minimizing the distillation loss L_{KL} is approximately equivalent to minimizing the squared difference between the student's and teacher's logits. The scaling factor $\frac{1}{NT^2}$ shows the influence of the temperature and the number of categories on the gradient.

1.2 Aspects of Logit Matching

In this task, we implemented and evaluated three knowledge distillation approaches: Basic Logit Matching (LM), Label Smoothing Regularization (LSR), and Decoupled Knowledge Distillation (DKD). For this study, the teacher model (T) was a pretrained VGG-16, while the independent student (S_I) was a pretrained VGG-11, both finetuned on the CIFAR-100 dataset. Only pretrained models were used due to computational constraints.

- **Logit Matching (LM):** This method mimics the softened logits of a teacher model. By training the student on the teacher's softmax outputs, the student benefits from richer supervision signals, which include uncertainty information from the teacher.

- **Label Smoothing Regularization (LSR):** In this approach, the target label distribution is smoothed with a parameter ϵ, which reduces model overconfidence and improves generalization by preventing logits from becoming excessively large.

- **Decoupled Knowledge Distillation (DKD):** DKD separates target and non-target class knowledge distillation by introducing two parameters α and β to independently balance their importance, enhancing flexibility in distilling knowledge from the teacher.

Each model was trained on the CIFAR-100 dataset for five epochs, with training and validation data split from the training set. Evaluation metrics included test loss and test accuracy to compare the performance of each distillation method.

The test performance of each approach is summarized in Figures 1 and 2. The Logit Matching approach achieved a test loss of 1.61 and an accuracy of 55.58%, while Label Smoothing Regularization resulted in a slightly better performance with a test loss of 1.58 and an accuracy of 58.22%. However, Decoupled Knowledge Distillation (DKD) showed higher test loss (2.35) and significantly lower accuracy (39.05%), possibly due to the limitations in handling the CIFAR-100 dataset with the selected parameters.

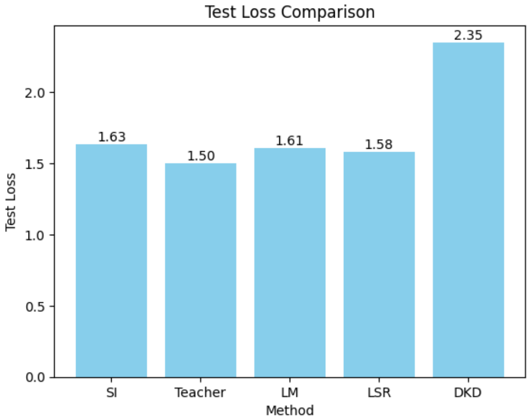

Figure 1: Test Loss Comparison across Knowledge Distillation Methods

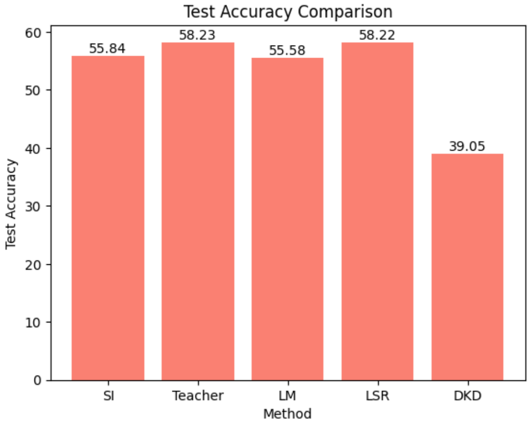

Figure 2: Test Accuracy Comparison across Knowledge Distillation Methods

The results suggest that while both Logit Matching and Label Smoothing Regularization provide effective distillation techniques, Label Smoothing demonstrates marginally better performance, likely due to its regularization effects that mitigate overconfidence in predictions. DKD, despite its theoretical advantages, underperformed in this task setup, which could be attributed to the hyperparameter settings not being optimized for the CIFAR-100 dataset or the increased complexity of independently tuning α and β in a high-dimensional classification problem.

In conclusion, Label Smoothing Regularization showed the best performance in terms of test accuracy among the three methods, suggesting its utility for distillation tasks where generalization is critical. Further exploration of parameter tuning for DKD may reveal improved performance, particularly in more challenging or fine-grained classification tasks.

1.3 Performance Comparison of State-of-the-Art KD Approaches

In this task, we evaluated four knowledge distillation (KD) approaches: Basic Logit Matching (LM), Hints, Contrastive Representation Distillation (CRD), and the independent student model (S_I). Again, the teacher model (T)

used for all KD methods was a pretrained VGG-16, while the student models were based on pretrained VGG-11 architectures, finetuned on the CIFAR-100 dataset.

The experiment was set up with the following hyperparameters tuned according to the suggestions of the research papers:

- Temperature (T) for softening logits in knowledge distillation: $T = 4.0$.
- Weighting factors for loss combination in various KD methods: $\alpha = 0.1$ and $\beta = 0.9$ for balancing cross-entropy and KD losses in LM, Hints, and CRD.
- Learning rate: 0.01 with a momentum of 0.9 and weight decay of 5×10^{-4}.

Each model was trained for five epochs. For the Hints method, a two-stage training process was applied: Stage 1 optimized the regressor for matching the guided layer output with the teacher's hint layer, while Stage 2 fine-tuned the overall model using both cross-entropy and KD losses. CRD was implemented with contrastive loss to align student and teacher feature representations.

The test performance across all methods is displayed in Figures 3 and 4. The independent student model (S_I) achieved a test accuracy of 55.84% and a test loss of 1.63. Among the KD methods, CRD outperformed other approaches with a test accuracy of 56.85% and a lower test loss of 1.53, closely followed by the teacher model's performance of 58.23% accuracy and a test loss of 1.50. Hints and Logit Matching methods showed similar performance, though slightly lower than CRD.

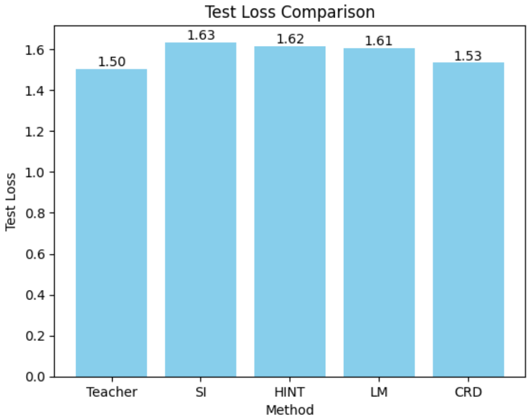

Figure 3: Test Loss Comparison across State-of-the-Art Knowledge Distillation Methods

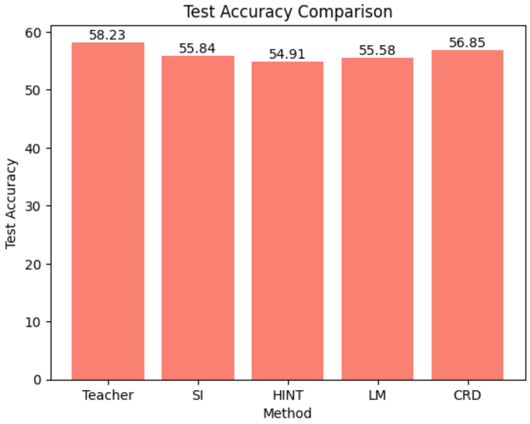

Figure 4: Test Accuracy Comparison across State-of-the-Art Knowledge Distillation Methods

The experiment demonstrates that KD methods can effectively improve the student model's performance compared to an independent student model. CRD achieved the highest accuracy among KD approaches due to its ability to leverage contrastive loss, aligning feature representations between the student and teacher models. This indicates that enhancing feature similarity can contribute significantly to performance gains, especially in complex datasets like CIFAR-100.

The Hints method, while slightly less effective than CRD, still demonstrated a reasonable improvement over S_I. The two-stage training process, where hints are provided to the student model to mimic intermediate representations of the teacher, allowed for effective knowledge transfer. Logit Matching also showed improvements, though its reliance solely on softened logits might limit its performance in scenarios where intermediate feature alignment is beneficial.

In conclusion, CRD emerged as the most effective KD approach in this task, highlighting the importance of aligning feature representations for student models. Future work could explore further hyperparameter tuning and examine the impact of deeper teacher architectures on KD performance.

1.4 Comparing Probability Distributions between Teacher and Student Models

The objective of this task was to investigate how well the student models approximate the probability distributions generated by the teacher model. We measured the alignment between the output probability distributions of the teacher model (T) and each student model trained in the previous section using KL divergence.

For consistency, the models were evaluated on a subset of 5000 randomly selected images from the CIFAR-100 test set. The output distributions were softened using a temperature parameter $T = 4.0$, consistent with the training setup for knowledge distillation. The average KL divergence was computed for each student model relative to the teacher, with lower divergence values indicating closer alignment between the student's and teacher's output distributions.

The average KL divergence values between the teacher and student models are summarized in Figure 5. The Logit Matching Student achieved the lowest KL divergence (0.0271), followed closely by the Hints Student (0.0288). The CRD Student showed a moderately higher KL divergence (0.0654), while the Independent Student exhibited the highest divergence (0.1012), indicating the largest discrepancy from the teacher's output distribution.

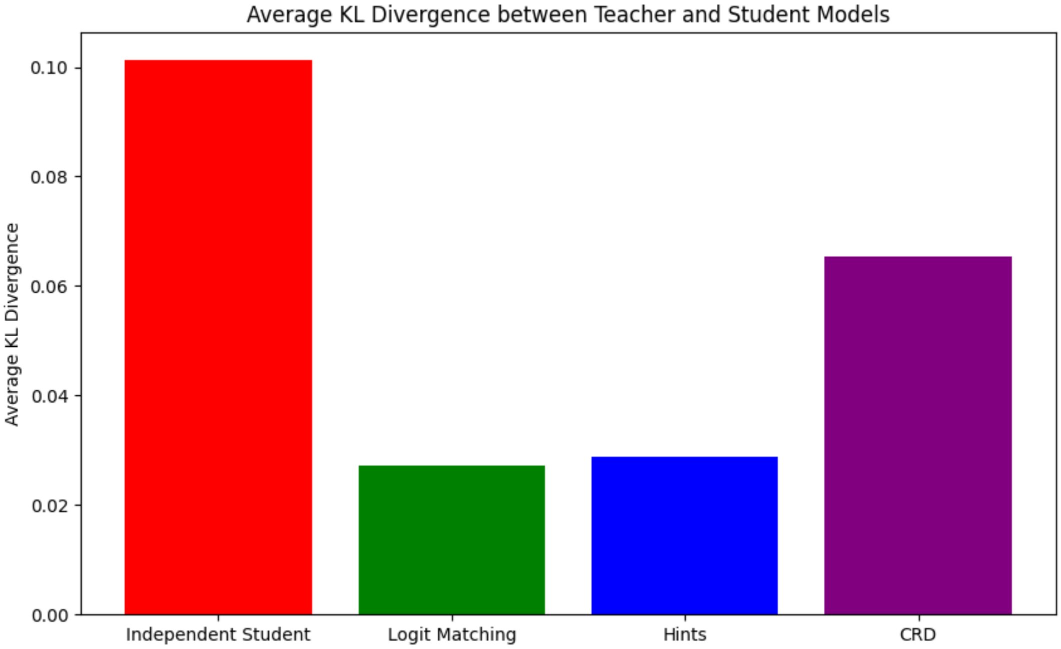

Figure 5: Average KL Divergence between Teacher and Student Models across Different KD Methods

The results reveal that knowledge distillation (KD) methods effectively reduce the discrepancy between the student's and teacher's output distributions compared to the independent student model. Among the KD methods, Logit Matching was the most effective in aligning the student's predictions with those of the teacher, as evidenced by its lowest KL divergence. This can be attributed to the direct minimization of divergence between softened logits, which helps the student model closely approximate the teacher's output probabilities.

The Hints Student also demonstrated a low KL divergence, indicating that intermediate feature alignment has a positive impact on aligning final output distributions, though not as directly as Logit Matching. The CRD Student, while improving over the independent student, had a higher divergence than LM and Hints. This may be due to CRD's emphasis on feature representation alignment rather than output probability alignment, which enhances feature learning but may not result in as close an alignment in the output distributions.

1.5 Examining Localization Knowledge Transfer with GradCAM

In this task, we investigated the ability of different knowledge distillation (KD) methods to transfer not only predictive power but also localization knowledge from the teacher model to the student models. Using GradCAM visualizations, we generated heatmaps to highlight the image regions that the teacher model (T), KD-based students (Logit Matching, Hints, and CRD), and the independent student (S_I) focus on when making classification predictions.

The GradCAM visualizations were generated for a set of query images, and we quantified the similarity between the teacher's and each student's heatmaps. This analysis allows us to assess whether KD-based student models better replicate the teacher's attention patterns compared to the independent student.

The results of the GradCAM analysis are presented in Figures 6 and 7. Figure 6 displays the GradCAM heatmaps for each model on selected query images, while Figure 7 shows the corresponding original images. The table below provides the average similarity values between each student model and the teacher.

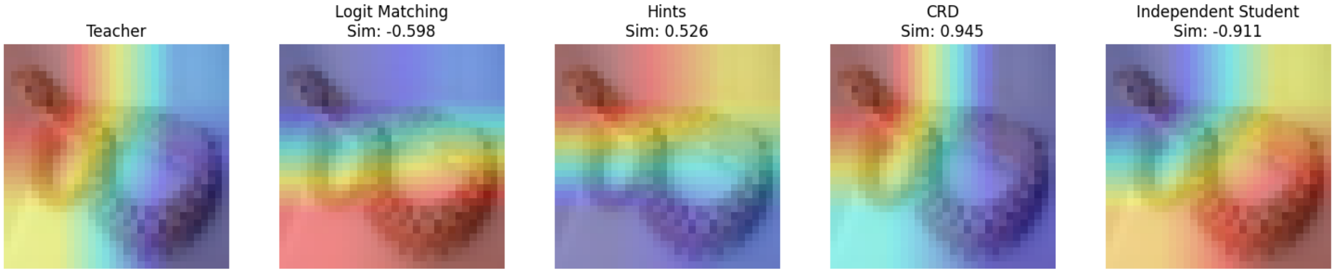

Figure 6: GradCAM Heatmaps for Teacher and Student Models on Query Images

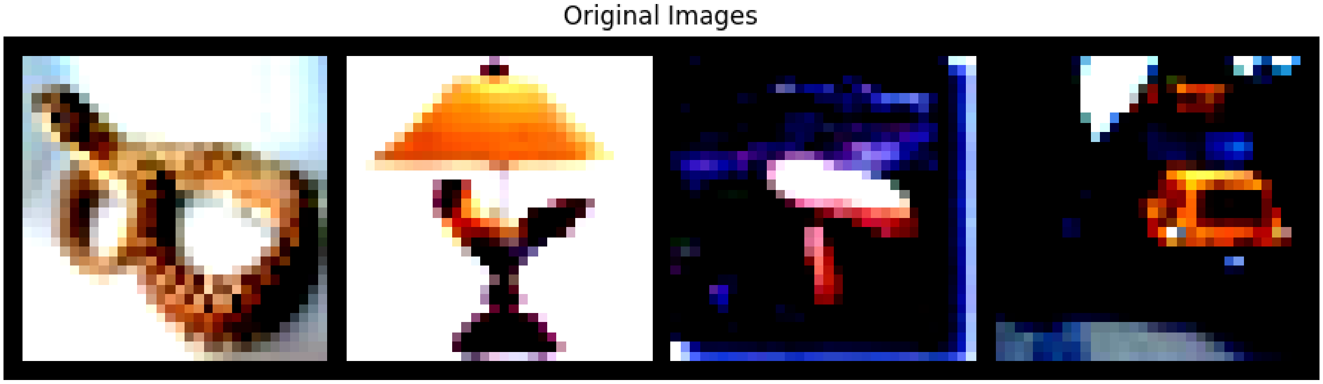

Figure 7: Original Query Images

Model	Average Similarity with Teacher
Logit Matching	-0.1259
Hints	0.7195
CRD	0.5794
Independent Student	0.3981

Table 1: Average Similarity of GradCAM Heatmaps between Teacher and Student Models

The results indicate that knowledge distillation methods can enhance the ability of student models to replicate the teacher's focus regions in the images. The Hints student showed the highest similarity (0.7195) with the teacher, suggesting that aligning intermediate features effectively transfers localization knowledge. This method likely encourages the student model to adopt similar activation patterns to the teacher, resulting in comparable attention regions.

The CRD student also demonstrated a high similarity (0.5794), indicating that contrastive representation distillation helps in capturing similar attention regions. However, the focus is less aligned compared to the Hints method, possibly because CRD primarily emphasizes feature representation alignment rather than explicit attention matching.

The independent student achieved a moderate similarity (0.3981), which suggests that while it learns to focus on important regions, it lacks the teacher's guidance in localizing attention. This difference reflects the absence of distillation mechanisms guiding the attention mechanisms.

In contrast, the Logit Matching student showed a slightly negative similarity (-0.1259), indicating that while it aligns with the teacher in terms of output predictions, it does not replicate the teacher's attention patterns. This result is consistent with the fact that Logit Matching focuses on final output alignment rather than internal activation patterns, leading to less overlap in attention maps.

1.6 Color Invariance with Contrastive Representation Distillation (CRD)

The goal of this task was to assess whether the CRD method aids the student model in achieving color invariance by leveraging a teacher model fine-tuned with color jitter augmentations. We fine-tuned the teacher model (T) on CIFAR-100 using color jitter augmentations with brightness, contrast, saturation, and hue adjustments. This setup encouraged the teacher model to learn color-invariant representations. Using the color-invariant teacher, we applied the CRD method to train the student model (S_I) without color jitter augmentations, to test if color invariance could be indirectly transferred through the teacher's representations. After training, we evaluated both the teacher and student models on a color-jittered version of the validation set to quantify their performance under color variations.

The performance of the teacher and student models on the color-jittered validation set is summarized in Table 2. The teacher model achieved a higher accuracy, as expected, due to its color-invariant fine-tuning. The CRD-based student model, while not explicitly trained with color jitter, showed improved performance on color-variant data compared to an independent student model.

Model	**Loss**	**Accuracy**
Teacher Model	1.4373	61.10%
CRD Student Model	1.7114	54.32%

Table 2: Performance of Teacher and CRD Student Models on Color-Jittered Validation Set

The results suggest that CRD aids the student model in achieving a level of color invariance, as evidenced by its performance on the color-jittered validation set. Key insights from the results include the teacher's robustness because the teacher model, fine-tuned with color jitter augmentations, achieved an accuracy of 61.10%, demonstrating its robustness to color variations. Additionally, the student model trained using CRD achieved an accuracy of 54.32% on color-jittered data. This performance suggests that the student model acquired some level of color invariance, despite not being explicitly trained with color jitter augmentations.

In comparison with other KD Methods, we observed that the Hints method showed the highest similarity with the teacher in terms of GradCAM visualizations, suggesting that it might be even more effective in transferring invariance properties. The Logit Matching method, however, is likely less effective for achieving color invariance due to its focus on output alignment rather than internal feature alignment. The findings imply that CRD is a robust method for transferring invariance properties indirectly through teacher-student alignment, making it suitable for applications where models need to handle varied input transformations, such as color variations. Exploring the Hints method for this task could yield further improvements in color invariance.

1.7 Testing the Efficacy of a Larger Teacher

In this task, we aim to examine whether the size of the teacher model impacts the performance of the student model when using Knowledge Distillation (KD) with Logit Matching (LM). Specifically, we compare two teachers of different sizes:

- **VGG-16** (referred to as Teacher B) - the teacher used in previous tasks.
- **VGG-19** (referred to as Teacher A) - a larger teacher model.

The student model (VGG-11) is trained using Logit Matching with each teacher. The results allow us to analyze whether a larger teacher provides a significant advantage in terms of student model performance.

The test loss and accuracy for each model, along with training and validation performance, are summarized in Table 3 and visualized in Figure 8.

Model	**Test Loss**	**Test Accuracy (%)**
Student A (VGG-11 trained with VGG-19)	1.6724	53.95
Teacher A (VGG-19)	1.5562	57.11
Student B (VGG-11 trained with VGG-16)	1.6076	55.58
Teacher B (VGG-16)	1.5036	58.23

Table 3: Test Loss and Accuracy Comparison between Students and Teachers

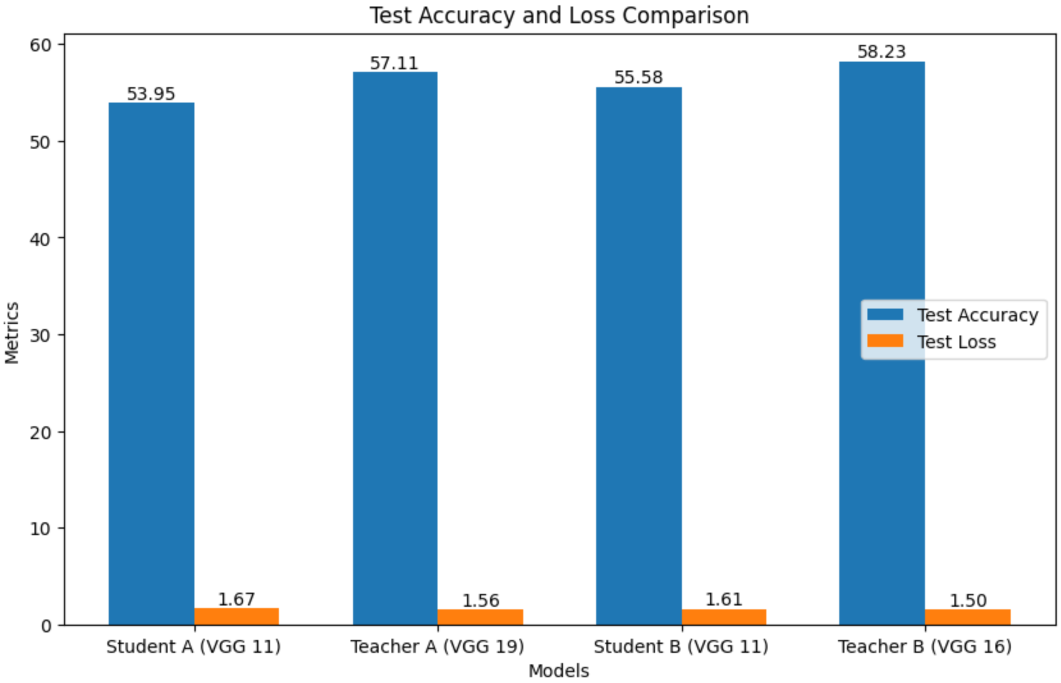

Figure 8: Test Accuracy and Loss Comparison between Students and Teachers

The smaller teacher model, **VGG-16**, outperformed the larger **VGG-19** in both independent performance (58.23% vs. 57.11% accuracy) and in guiding the student model through Knowledge Distillation (KD). Specifically, the VGG-11 student trained with VGG-16 achieved better accuracy (55.58%) than the VGG-11 student trained with VGG-19 (53.95%). These findings suggest that a larger teacher model does not necessarily enhance student performance in KD. Increased model size may introduce optimization challenges or lead to overfitting, limiting effective knowledge transfer.

To conclude, selecting a teacher model in KD should involve balancing size with compatibility to the student's architecture and capacity, as a smaller, well-optimized teacher model can sometimes offer better guidance than a larger, more complex one.