Exploring the Impact of Hyperparameters on the Effectiveness of Knowledge Distillation in Multiclass Image Classification

Guru Prakash Ram Balamurugan
Student Number 220861931
MSc. Artificial Intelligence, QMUL
g.balamurugan@se22.qmul.ac.uk
Project Supervisor: Charalampos Saitis
c.saitis@qmul.ac.uk

***Abstract*—Knowledge distillation, where a smaller student model learns from a pre-trained larger teacher model, offers a pathway to condense the knowledge of substantial models into more deployable ones. This research investigates the critical roles of hyperparameters, specifically the temperature parameter in the softmax function and the weighting of the distillation loss, in the multiclass image classification task using the CIFAR-10 dataset. Employing ResNet-50 as the teacher model, and ResNet-18 and MobileNetv3-small as student models, the study explored optimal hyperparameters for maximizing accuracy. The results unveiled a nuanced interplay between hyperparameters and model architecture, with ResNet-18 demonstrating stability across distillation loss weights (89.94% to 90.81%) and MobileNet exhibiting sensitivity, with a test accuracy range of 80.54% to 85.19%). Furthermore, the findings revealed that the distilled models were more robust to adversarial attacks than their counterparts trained on hard targets, emphasizing the multifaceted benefits of knowledge distillation. These insights highlight the importance of tailored hyperparameter tuning and provide valuable guidelines for harnessing the full potential of knowledge distillation, contributing significantly to the field and offering practical applications for enhanced model performance and security.**

***Index Terms*—knowledge distillation, hyperparameter tuning, softmax temperature, distillation loss weighting**

I. Introduction

The deep learning realm has produced an array of architectures tailor-made for diverse, intricate tasks, especially within computer vision LeCun et al. (2015). Residual Networks (ResNet), renowned for their skip connections, stand out for simplifying the training of profoundly deep networks, subsequently amplifying accuracy in image classification He et al. (2016). However, such profound architectures, while accurate, grapple with computational constraints, making them less ideal for environments with limited resources, such as mobile devices Howard et al. (2017).

"Knowledge distillation" emerges as a beacon in this scenario, enabling a smaller student model to mimic its more extensive teacher model, thus compacting the latter's expertise Hinton et al. (2015). The mechanics of this approach are heavily tethered to the temperature-scaled softmax function and the precise weighting of the distillation loss – two components that wield considerable influence over the student model's efficacy Polino et al. (2018).

We aim to explore the intricate effects of these hyperparameters within knowledge distillation, focusing on image classification using the CIFAR-10 dataset with the ResNet and MobileNetv3 architectures. Beyond presenting empirical insights, this research aspires to carve out pragmatic guidelines to refine knowledge distillation processes, making pioneering models more amenable to real-world applications.

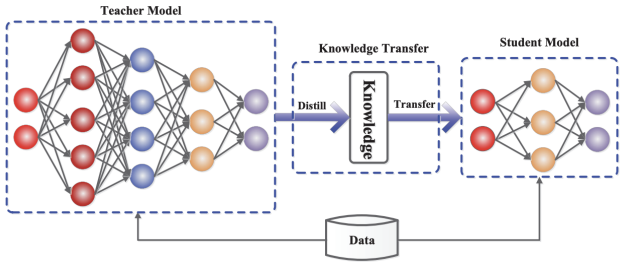

Fig. 1: Illustration of the response-based knowledge distillation process Gou et al. (2021).

In complementing this primary objective, the research acknowledges the imperative of evaluating distilled models beyond sheer accuracy. Given the escalating concerns over neural network susceptibility to adversarial onslaughts, the robustness of these models demands scrutiny Szegedy et al. (2013). Enter the Fast Gradient Sign Method (FGSM) – a notable technique in adversarial attacks – serving as our chosen evaluative metric Goodfellow et al. (2014). Through this lens, we aim to discern whether distilled models possess enhanced defenses against adversarial threats relative to their counterparts trained on conventional hard targets.

In summary, this study sheds light on the complex interplay of hyperparameters in knowledge distillation, with an emphasis on understanding their impact on student model performance and robustness. The findings offer valuable insights and practical guidelines, contributing significantly to the field

of deep learning and opening new paths for exploration and innovation.

II. Background

A. Knowledge Distillation

Knowledge distillation (KD) is a technique in which a smaller student model learns from a larger teacher model's soft outputs, often achieving higher performance than independent training Cho & Hariharan (2019).

1) Key Ideas:

a) Soft Probabilities:: The soft probabilities from the teacher model reveal underlying relationships between classes. Computed using the softmax function with a temperature parameter T, the soft probabilities are defined as:

$$P^i_{\text{teacher}} = \frac{\exp\left(\frac{z_i}{T}\right)}{\sum_j \exp\left(\frac{z_j}{T}\right)}, \tag{1}$$

where P^i_{teacher} is the probability for class i, and z_i is the logit for that class.

b) Student Mimicking:: The student model learns by minimizing a combined loss function of hard targets (original labels) and soft targets (teacher's outputs), with the soft target loss computed using the Kullback-Leibler (KL) divergence:

$$\mathcal{L}_{\text{soft}} = T^2 \cdot \text{KL}\left(P_{\text{teacher}} || P_{\text{student}}\right), \tag{2}$$

$$\mathcal{L} = \alpha \mathcal{L}_{\text{hard}} + (1 - \alpha)\mathcal{L}_{\text{soft}}, \tag{3}$$

where α balances hard and soft targets.

c) Improved Generalization:: KD often enhances student model generalization and robustness, leading to better performance on unseen data.

d) Conclusion: KD leverages nuanced information from the teacher model to guide student training, resulting in improved performance. The utilization of temperature scaling and the combination of hard and soft loss components are essential. KD principles continue to be an active research area with various applications Cho & Hariharan (2019).

B. Fast Gradient Sign Method (FGSM)

The Fast Gradient Sign Method (FGSM) is a white-box adversarial attack used to assess neural network robustness by applying small perturbations to input images, leading to misclassification without significant visual alteration (see Figure 2) Goodfellow et al. (2014).

1) Key Ideas:

a) Adversarial Perturbation:: FGSM computes a perturbation by taking the gradient of the loss function with respect to the input image, and adjusting in the direction that increases the loss:

$$\Delta x = \epsilon \cdot \text{sign}\left(\nabla_x \mathcal{L}(\theta, x, y)\right), \tag{4}$$

where ϵ is a small constant, and $\nabla_x \mathcal{L}(\theta, x, y)$ represents the gradient of the loss function Goodfellow et al. (2014).

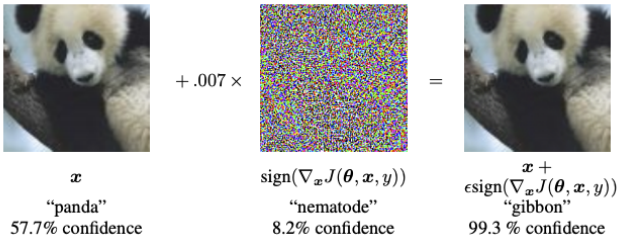

Fig. 2: An adversarial attack on an image of a panda, where small perturbations are added to the original image, leading the model to misclassify it as a gibbon with high confidence Goodfellow et al. (2014).

b) Generation of Adversarial Example:: The adversarial example is formed by adding the perturbation to the original image:

$$x_{\text{adv}} = x + \Delta x, \tag{5}$$

resulting in a visually similar image that may cause incorrect classification Goodfellow et al. (2014).

c) Model Vulnerability and Robustness Evaluation:: FGSM uncovers vulnerabilities in even state-of-the-art models, emphasizing non-robust features that lead to incorrect predictions. It serves as a tool to evaluate and improve model resilience against adversarial manipulation Goodfellow et al. (2014).

d) Conclusion: FGSM illuminates the susceptibility of neural networks to adversarial perturbations and has spurred extensive research into adversarial attacks and defenses. Its continued exploration is essential for developing more secure and reliable machine learning systems Goodfellow et al. (2014).

III. Related Work

Knowledge distillation (KD), where a student model is trained to replicate the behavior of a teacher model, hinges on specific hyperparameters like temperature scaling and distillation loss weighting. This section reviews literature related to our study's emphasis on these hyperparameters, multiclass image classification, and robustness to adversarial attacks.

A. Hyperparameter Tuning in Knowledge Distillation

While KD has been widely explored, the specific aspects of temperature and distillation loss tuning have remained relatively underexplored. Liu et al. (2022) introduced Meta Knowledge Distillation, adapting temperature during training, and offering insights into its tuning. Li et al. (2023) built upon this by dynamically controlling temperature through Curriculum Temperature for Knowledge Distillation. These works align with our investigation into optimal hyperparameters for KD.

B. Adherence to Conventional Hyperparameters in Knowledge Distillation

In some studies, the hyperparameters for knowledge distillation, specifically the distillation loss weight and temperature scaling, are selected based on established practices. For example, Cho et al. Cho & Hariharan (2019) conducted experiments on CIFAR10 and ImageNet, adhering to popular choices for hyperparameters. They set the temperature parameter $\tau = 4$, $\alpha = 0.9$, and $\beta = 1000$ for attention transfer, aligning with prior works Zagoruyko & Komodakis (2016), Hinton et al. (2015).

Our research, conversely, adopts a more experimental approach by investigating a range of distillation loss weights and temperature values. We observed that optimal hyperparameters are not universally applicable but are specific to the student model's architecture and characteristics. By demonstrating that a fixed distillation loss weight and temperature may not always yield optimal performance, our findings challenge the assumption of universally optimal values and emphasize the importance of model-specific tuning. This insight underscores the multifaceted nature of knowledge distillation and the need for careful empirical validation in hyperparameter selection.

C. Knowledge Distillation and Model Robustness

The robustness of KD models to adversarial attacks has been explored by Zhao et al. (2023) and Papernot et al. (2016). These studies contextualize our evaluation of distilled models' resilience to adversarial threats.

D. Challenges and Complexities in Knowledge Distillation

Stanton et al. (2021) identified challenges in KD, emphasizing that a close match between teacher and student models does not always improve generalization. These insights inform our exploration of the complex relationships between architecture, hyperparameters, and performance.

E. Conclusion

The related work highlights the multifaceted nature of KD and the critical role of hyperparameters. It also reveals the existing challenges in model robustness. Building upon these insights, our study aims to empirically determine optimal temperature and distillation loss weight in multiclass image classification, contributing to the broader understanding of KD.

IV. Methods

A. Environment and Tools

The experiments were conducted using Google Colab, leveraging the powerful NVIDIA Tesla V100 GPU to accelerate the training process. Specific packages, such as tensorboardX, were installed to facilitate the implementation and enable real-time monitoring of the training process

B. Dataset

The study utilized the CIFAR-10 dataset, comprising 60,000 32x32 color images equally distributed across 10 different classes Krizhevsky et al. (2009). The dataset was divided into 50,000 training images and 10,000 test images.

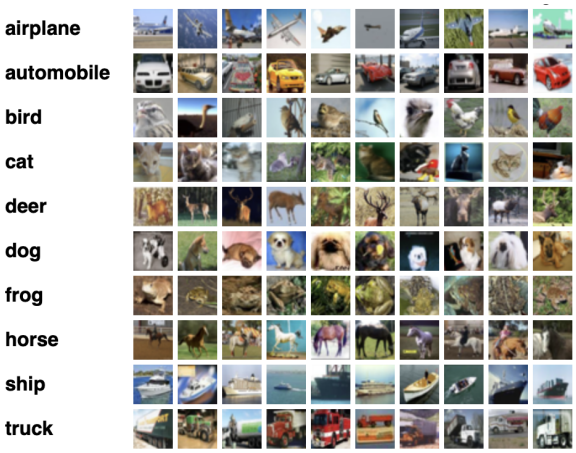

Fig. 3: CIFAR-10 dataset sample images Krizhevsky (n.d.).

C. Data Preprocessing

1) Reproducibility: To ensure consistent results, random seeds were set for all random number generators, including PyTorch and NumPy.

2) Computing Dataset Statistics: The mean and standard deviation of the CIFAR-10 dataset were computed and used for image normalization.

3) Transformations: Images underwent several transformations to augment the dataset and improve model generalization. These transformations included:

- Padding: Extending the image borders by 4 pixels.
- Random Horizontal Flipping: Flipping the image horizontally with a 50% probability.
- Random Rotation: Rotating the image by up to 15 degrees.
- Random Cropping: Cropping the image to 32x32 pixels.
- Normalization: Normalizing the image using the mean and standard deviation computed from the training set.

4) Data Splitting and Loading: The full training dataset was randomly split into training and validation sets using a 90:10 ratio. Data loaders were created with a batch size of 300 and 2 worker threads for efficient data loading.

D. Model Architecture

1) Teacher Model: ResNet-50:

a) ResNet-50 (Teacher Model): The teacher model was built using the ResNet-50 architecture, a deeper variant of the Residual Network with 50 layers, and was adapted to suit the CIFAR-10 dataset. The architecture included the following components:

- Convolutional layers with a kernel size of 3×3, specifically modifying the first layer to have a kernel size of 3×3, a stride of 1, and padding of 1, without bias.
- Batch normalization to stabilize and accelerate the learning process.
- ReLU activation functions to introduce non-linearity without affecting the receptive fields of the convolution layer.

- A fully connected layer with 10 output units to correspond to the 10 classes of the CIFAR-10 dataset.

These modifications ensured that the teacher model was tailored to the specifics of the CIFAR-10 dataset, providing an effective basis for the knowledge distillation process.

2) Student Models: ResNet-18 and MobileNetV3-small:

a) ResNet-18: ResNet-18 is a deep residual network with 18 layers that is known for its capability to train very deep networks by using shortcut connections. For the CIFAR-10 classification task, the architecture was modified to suit the dataset. The first convolutional layer was altered to have a kernel size of 3×3, a stride of 1, and padding of 1, without bias. The fully connected layer was also modified to match the number of CIFAR-10 classes.

b) MobileNetV3-small: MobileNetV3-small is a lightweight and efficient architecture designed for mobile and embedded vision applications. It employs depthwise separable convolutions that reduce the computational burden without compromising accuracy. Similar to ResNet-18, modifications were made to fit the CIFAR-10 dataset. The first convolutional layer was adjusted to have a kernel size of 3×3, a stride of 1, and padding of 1, without bias, and the classifier was adapted to the number of CIFAR-10 classes.

E. Training Process

1) Training with Hard Targets: The training process was carried out using the Cross-Entropy Loss function and optimized using the Adam optimizer with an initial learning rate of 0.001 and a weight decay of 1×10^{-4}. The learning rate was managed using PyTorch's ReduceLROnPlateau scheduler, which monitors the validation loss and reduces the learning rate by a factor of 0.2 if no improvement is observed for 5 consecutive epochs (patience = 5). Additionally, an early stopping method with a patience of 10 was implemented to define the number of epochs based on the convergence of the model, ensuring a robust and systematic approach to multiclass image classification. The batch size was set to 300, and the model was trained with hard targets.

2) Training with Soft Targets (Knowledge Distillation): Knowledge distillation was implemented by training the student models using soft targets from the teacher model. In the experimentation phase, several hyperparameters were strategically tuned to optimize the model's performance. The temperature parameter, crucial for the softmax function, was varied across the values [1,2,4,6,8,10,15,20], allowing a detailed investigation of its impact on the student model's learning. Distillation loss, a combination of Cross-Entropy Loss with both soft targets and hard targets, was initially set at 0.9, a value guided by the findings of Hinton et al. (2015) and Zagoruyko & Komodakis (2016), who demonstrated the effectiveness of this specific value in their study. Using the learning rate scheduler employed for training with hard targets, a series of experiments were first conducted to identify the best-performing temperature from the aforementioned range. Subsequently, further analysis was carried out by experimenting with different distillation loss values ranging from 0.0 to 1.0, in increments such as 0.1, 0.3, 0.5, 0.7, and 0.9, to scrutinize the influence of this parameter on the student model's multiclass image classification performance.

V. Results

The results section begins by presenting the baseline test accuracies for the teacher and student models trained under normal conditions without knowledge distillation. Table I summarizes these baseline accuracies. The teacher model, ResNet-50, achieves the highest test accuracy of 91.36%. ResNet-18, a student model with a similar architecture but reduced complexity, follows closely with a test accuracy of 88.80%. MobileNet, another student model with a more compact architecture, registers a test accuracy of 81.35%. These baseline results provide a point of reference for the subsequent analysis, where the impact of knowledge distillation and hyperparameter tuning on the student models is explored in detail. The following subsections present the outcomes of various experiments. These experiments were conducted to investigate how temperature scaling and distillation loss weighting influence both the performance and robustness to adversarial attacks of the student models.

Model Type	Test Accuracy (%)	Epsilon = 0.01 (%)	Epsilon = 0.05 (%)	Epsilon = 0.1 (%)
ResNet-50	91.36	71.68	23.73	11.14
ResNet-18	88.80	71.09	19.97	5.60
MobileNet	81.35	61.25	12.19	3.07

TABLE I: Test accuracy and robustness to FGSM attacks for Resnet-50, Resnet-18, and MobileNet under normal training conditions.

A. Analysis of Test Accuracy and Robustness to FGSM Attacks in Student Models

1) MobileNet: The optimal combination of hyperparameters (Temperature = 4, Distillation Loss Weight = 0.3) achieves a test accuracy of 85.19%, exceeding the hard target model by an accuracy boost of 3.84%.

Similarly, the MobileNet model trained with knowledge distillation achieves higher robustness to FGSM attacks across all tested epsilon values, with an accuracy boost ranging from 1.73% to 3.43% compared to the model trained on hard targets.

Training Type	Hard Targets (%)	Optimal Combination (%)	Accuracy Boost (%)
Test Accuracy	81.35	85.19	3.84

TABLE II: Test Accuracy Comparison for MobileNet

Epsilon	Hard Targets (%)	Optimal Combination (%)	Accuracy Boost (%)
0.01	61.25	64.68	3.43
0.05	12.19	14.10	1.91
0.1	3.07	4.80	1.73

TABLE III: Robustness to FGSM Attacks for MobileNet

2) Resnet-18: The optimal combination of hyperparameters (Temperature = 6, Distillation Loss Weight = 0.9) achieves a test accuracy of 90.72%, exceeding the hard target model by an accuracy boost of 1.92%. Similarly, the ResNet-18 model trained with knowledge distillation achieves higher robustness to FGSM attacks across all tested epsilon values, with an accuracy boost ranging from 1.85% to 3.48% compared to the model trained on hard targets.

Training Type	Hard Targets (%)	Optimal Combination (%)	Accuracy Boost (%)
Test Accuracy	88.80	90.72	1.92

TABLE IV: Test Accuracy Comparison for ResNet-18

Epsilon	Hard Targets (%)	Optimal Combination (%)	Accuracy Boost (%)
0.01	71.09	73.44	2.35
0.05	19.97	23.45	3.48
0.1	5.60	7.45	1.85

TABLE V: Robustness to FGSM Attacks for ResNet-18

B. Analysis of Distillation Loss Weight Impact on Student Model's Test Accuracy

a) Baseline:

- Distillation Loss Weight = 0: This baseline represents the scenario where the student model is trained solely on the hard targets without any contribution from the soft targets. The test accuracy at this baseline (indicated by the red dashed line) serves as a reference point to understand how introducing soft targets influences the model's performance.
- Distillation Loss Weight = 1: Conversely, the second baseline with a distillation loss weight of 1 (indicated by the green dashed line) represents the extreme where the student model is trained entirely on the soft targets. This provides a contrasting view of the impact of soft targets on model performance.

1) MobileNet: In Figure 4, the test accuracy of the MobileNet model is analyzed across various distillation loss weights, with a fixed temperature value of 4. The plot provides insights into how MobileNet's performance is influenced by the distillation loss weight in the knowledge distillation process.

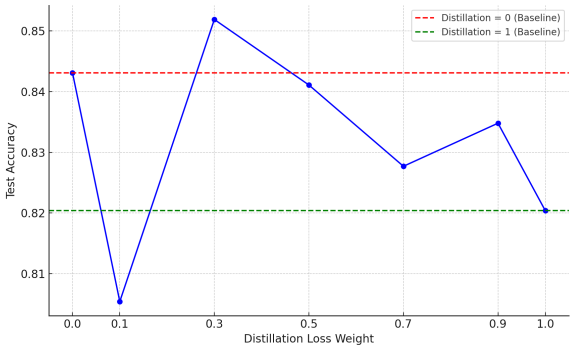

Fig. 4: Test Accuracy vs. Distillation Loss Weight for MobileNet with fixed Temperature = 4. Baselines with Distillation Loss Weights of 0 and 1 are indicated by dashed lines.

a) Performance Across Distillation Loss Values:

- **Optimal Range**: The plot reveals an optimal range of distillation loss weights between 0.1 and 0.7 where the MobileNet model achieves higher test accuracy compared to both baselines. Particularly, at a distillation loss weight of 0.3, a peak test accuracy of 85.19% is observed, providing a beneficial balance between hard and soft targets.
- **Sensitivity to Changes**: A careful analysis of the plot shows that small changes in the distillation loss weight within the optimal range lead to significant variations in test accuracy, indicating a sensitive dependence on this hyperparameter. This underscores the importance of fine-tuning the distillation loss weight for optimal performance.
- **Declining Accuracy**: Outside the optimal range, especially at higher distillation loss weights close to 1, the test accuracy appears to decline to 82.04% approaching the performance at the baseline with a distillation loss weight of 1.
- **Interpretation**: The findings illustrate the nuanced role of the distillation loss weight in knowledge distillation for MobileNet. The optimal balance between hard and soft targets enhances the model's performance, while extremes in either direction lead to suboptimal results. This highlights the importance of carefully selecting the distillation loss weight in the knowledge distillation process.

2) ResNet-18: Figure 5 presents the test accuracy of the ResNet-18 model as a function of distillation loss weights, with the temperature fixed at 6. The plot offers distinct observations specific to the behavior of the ResNet-18 model in the knowledge distillation process.

a) Performance Insights::

- **Stability in Test Accuracy**: ResNet-18 exhibits a stable performance across different distillation loss weights, with test accuracy values staying within a narrow range

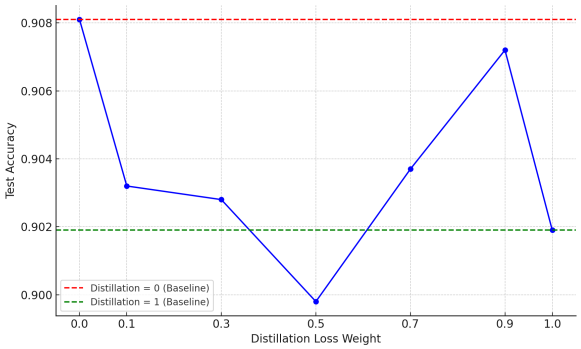

Fig. 5: Test Accuracy vs. Distillation Loss Weight for ResNet-18 with fixed Temperature = 6. Baselines with Distillation Loss Weights of 0 and 1 are indicated by dashed lines.

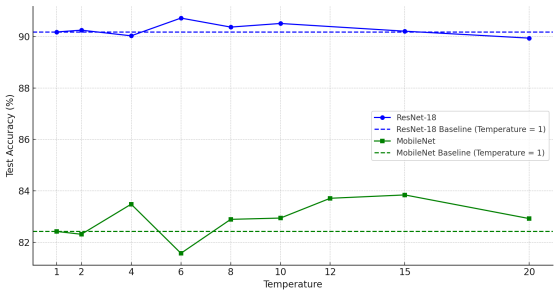

Fig. 6: Test Accuracy vs. Temperature for ResNet-18 and MobileNet with fixed Distillation Loss Weight = 0.9. The baseline with Temperature = 1 is indicated by a dashed line.

of 89.98% to 90.81%. This reflects a degree of robustness and a unique sensitivity pattern where the model's accuracy is consistently high.

- **Comparison with Baselines**: The test accuracy varies intricately across various distillation loss weights, surpassing, approximating, or falling below the baselines at different points. This demonstrates a complex balance between the influence of hard and soft targets, highlighting the model's distinct behavior.
- **Implications and Interpretation**: The stability and intricate balance observed in the plot indicate the nuanced role of the distillation loss weight for ResNet-18. The findings emphasize the importance of understanding and tuning this hyperparameter according to the specific characteristics of the student model, reflecting ResNet-18's unique learning dynamics in the knowledge distillation process.

C. Analysis of Temperature Impact on Student Model's Test Accuracy

In Figure 6, the test accuracy of the ResNet-18 and MobileNet models is analyzed across various temperature values, with a fixed distillation loss weight of 0.9. This plot serves to examine how the temperature parameter in the softmax function influences the student models' performance in the knowledge distillation process.

1) MobileNet:

a) Performance Insights:

- **Baseline Performance (Temperature = 1)**: The baseline test accuracy at a temperature of 1 is 82.42%. This serves as a reference to understand the effect of varying temperature on MobileNet's performance.
- **Sensitivity to Temperature Changes**: MobileNet's test accuracy varies from a minimum of 81.57% to a maximum of 83.84%, indicating a sensitivity to changes in temperature.
- **Optimal Performance (Temperature = 4)**: A peak in test accuracy is observed at a temperature of 4, achieving 83.48%. This temperature value appears to offer the most effective balance in the knowledge distillation process.
- **Declining Accuracy**: Outside the optimal range, especially at higher temperatures, the test accuracy declines, reflecting the nuanced role of temperature in MobileNet's performance.

2) ResNet-18:

a) Performance Insights::

- **Baseline Performance (Temperature = 1)**: The baseline test accuracy at a temperature of 1 is 90.18%. This serves as a reference to understand the effect of varying temperature on ResNet-18's performance.
- **Stable Performance Across Temperatures**: ResNet-18's test accuracy varies within a narrow range of 89.94% to 90.72%, indicating stable performance and less sensitivity to changes in temperature.
- **Optimal Performance**: The plot shows a gradual increase in test accuracy with temperature, reaching a peak at the higher end. This may indicate an optimal temperature that maximizes test accuracy for ResNet-18.
- **Implications and Interpretation**: The stability and gradual increase in test accuracy reflect the unique behavior of ResNet-18 in response to temperature changes.

D. Comparative Analysis of Temperature and Distillation Loss Weight Impact on ResNet-18 and MobileNet in Knowledge Distillation

The examination of both ResNet-18 and MobileNet models, as depicted in Figures 7 and 8, illustrates the MobileNet and Resnet-18 model's peak accuracy at specific hyperparameters, requiring tailored softening of logits. It underscores the complexity of knowledge distillation and the intricate relationships between model architecture, hyperparameters, and performance.

1) Model Architecture and Distillation Interaction: :

a) Distillation Loss Weight: ResNet-18's stable test accuracy across distillation loss weights (89.94% to 90.72%) may reflect its architectural compatibility with the ResNet-50 teacher model. In contrast, MobileNet, with a test accuracy

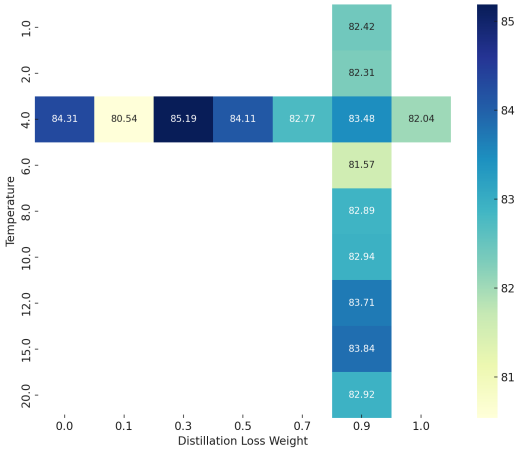

Fig. 7: MobileNet Test Accuracy for Different Temperatures and Distillation Loss Weights

range of 81.57% to 83.84%, exhibits a distinct peak, emphasizing the interplay between its unique structure and the distillation process.

b) Temperature Scaling: MobileNet's performance reveals an optimal temperature range of 2 to 6, while ResNet-18's test accuracy varies within a narrower range of 89.94% to 90.72%, indicating stability across temperatures. These observations highlight the nuanced role of temperature scaling in the performance of different student models.

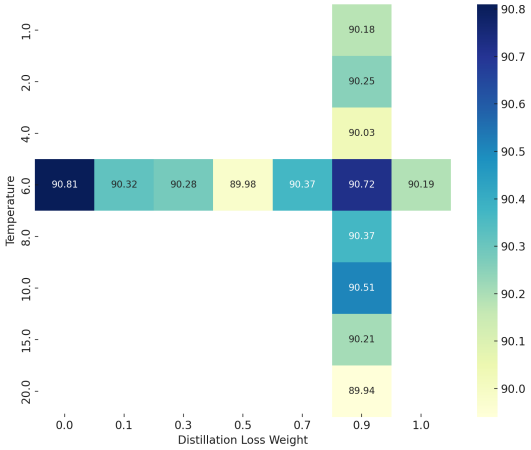

Fig. 8: ResNet-18 Test Accuracy for Different Temperatures and Distillation Loss Weights

2) Size of Models and Performance Sensitivity: :

a) Distillation Loss Weight: : MobileNet's lightweight design translates to a more pronounced sensitivity to distillation loss weight, as observed in the clear peak in test accuracy. ResNet-18, though more substantial, demonstrates a nuanced and stable response.

b) Temperature Scaling: : ResNet-18's stable performance across different temperatures as shown in Figure 8 may be influenced by its architectural similarity to the teacher model. MobileNet's sensitivity to temperature changes reflects its architectural distinctiveness and offers insights into how model size and complexity interact with temperature in the knowledge distillation process.

These observations emphasize the multifaceted nature of knowledge distillation, where architectural compatibility between teacher and student models, the inherent characteristics of the models, and the choice of hyperparameters all interplay.The differences also reflect the sensitivity of the distillation process to the dataset and training configuration. They emphasize the tailored nature of knowledge distillation, where each student model's response to hyperparameters is shaped by its unique characteristics.

E. Discussion and Implications

The findings of this study illuminate critical aspects of knowledge distillation, particularly emphasizing the nuanced roles played by temperature scaling and distillation loss weight in student model performance. Several key implications arise from these results.

a) Model Specificity: The distinct responses of ResNet-18 and MobileNet to changes in temperature and distillation loss weight underscore the need for tailored hyperparameter tuning. This observation challenges a one-size-fits-all approach and highlights the importance of understanding the specific characteristics of individual models.

b) Optimization of Student Model Performance: The research demonstrates that through careful adjustment of temperature and distillation loss weight, student models can achieve enhanced test accuracy and robustness to FGSM attacks. These findings provide a valuable guide for practitioners aiming to optimize model performance without significantly increasing complexity.

c) Insights into Model Architecture: The study also offers insights into how architectural differences between teacher and student models influence the knowledge distillation process. The performance variations between ResNet-18 and MobileNet reveal how architectural compatibility, inherent characteristics, and hyperparameters interplay, shedding light on the complex dynamics of knowledge distillation.

d) Robustness to Adversarial Attacks: The observed increase in robustness to FGSM attacks in models trained with knowledge distillation is a promising result. It suggests potential avenues for further research into the use of knowledge distillation as a tool for enhancing model security and resilience.

e) Baseline Comparisons: The detailed comparison between hard target training and knowledge distillation elucidates the added value of soft targets. This comparative analysis serves as a foundation for understanding when and how to employ knowledge distillation in various applications.

In summary, this research contributes to the evolving field of knowledge distillation by providing empirical evidence on the critical influence of temperature and distillation loss weight. These insights have broad applications, from optimizing performance in resource-constrained environments to enhancing model robustness. The study's results serve as a stepping stone for future research and practical applications, enriching the understanding of knowledge distillation and opening new paths for exploration and innovation.

F. Limitations

1) Dataset and Hyperparameter Constraints: This study faced several limitations that may have influenced the findings. Primarily, the experiments were constrained to the CIFAR-10 dataset. While this dataset is widely used, it may not capture all the complexities found in diverse real-world scenarios. Incorporating broader selections of datasets, such as ImageNet and CIFAR-100, would likely provide more general insights into the knowledge distillation process.

The chosen range of temperature and distillation loss weights was limited, possibly restricting the generalizability of the results. A more extensive exploration of these hyperparameters could uncover novel trends and relationships, enhancing the robustness of the findings.

2) Computational Constraints and Model Scope: GPU constraints significantly affected the extent of experiments. The usage of Google Colab, while providing necessary computational resources, incurred substantial costs that limited the ability to conduct more extensive training and trials. Furthermore, the focus on specific models like ResNet and MobileNet may not represent the full spectrum of architectures. A broader investigation encompassing various architectures could offer a more comprehensive understanding.

Lastly, the robustness evaluation concentrated solely on FGSM attacks. This focus leaves the model's response to other types of adversarial attacks unexplored, potentially overlooking crucial aspects of model vulnerability.

G. Future Work

1) Dataset and Model Expansion: Future research should consider employing additional datasets to obtain a more comprehensive understanding of the knowledge distillation process. ImageNet and CIFAR-100, for instance, would provide more varied and complex data, enhancing the study's applicability to real-world tasks.

Exploring various teacher and student models could lead to fresh insights into the knowledge distillation process and its dependencies on model architecture. Such exploration may uncover model-specific behaviors and responses, enriching the overall understanding of knowledge distillation.

2) Hyperparameter Exploration and Robustness Analysis: A more extensive search over temperature and distillation loss weight values would contribute to more generalized conclusions, allowing for the identification of optimal hyperparameters across diverse scenarios.

Moreover, an inclusion of different adversarial attacks, beyond FGSM, could provide a comprehensive view of model robustness. This expanded analysis would offer valuable insights into the models' resilience to various attack strategies, informing the development of more secure and reliable systems.

3) Scalability and Ethical Considerations: Assessment of scalability to more complex real-world scenarios and investigation of other knowledge distillation methods would further enrich the field. An analysis of potential biases, ethical implications, and the impact of knowledge distillation on fairness and transparency in model predictions could also be vital avenues for future exploration.

In summary, the limitations and future directions identified set the stage for continued investigations, enriching the understanding of knowledge distillation, and opening new paths for exploration, innovation, and responsible AI development.

VI. Conclusion

This dissertation presented a comprehensive investigation into the influence of key hyperparameters, specifically the temperature parameter in the softmax function and the distillation loss weighting, on a student model's performance in multiclass image classification through basic knowledge distillation. Utilizing CIFAR-10 as the dataset, ResNet-50 was trained as the teacher model, while ResNet-18 and MobileNetV3-small served as student models.

The experimental results revealed the critical role of carefully selected temperature values and distillation loss weights in enhancing the performance of student models. In the context of our experiments, the temperature values that yielded the best results were 6 for ResNet-18 and 4 for MobileNetV3, demonstrating that the effectiveness of knowledge distillation is indeed influenced by model architecture. The analysis further extended to robustness against FGSM adversarial attacks, highlighting the potential vulnerabilities and strengths of different configurations.

However, the study was not without limitations, including constraints related to the dataset, GPU resources, and the scope of hyperparameter tuning. These limitations prompted future work recommendations, such as the exploration of additional datasets, a broader range of hyperparameters, different teacher and student models, and ethical considerations.

The findings of this research contribute to the growing body of knowledge in the field of knowledge distillation, providing valuable insights into hyperparameter optimization and the interplay between teacher and student models. The study's methodology, including data preprocessing, model architectures, training procedures, and evaluation metrics, lays the groundwork for further investigations and applications.

In summary, this dissertation serves as a vital step towards a deeper understanding of knowledge distillation, fostering further innovation and exploration in multiclass image classification. The experimental design, results, and implications set a precedent for future research, with the potential to impact both academia and industry alike.

VII. Acknowledgments

I extend heartfelt gratitude to my supervisor, Dr. Charalampos Saitis, for his invaluable guidance and encouragement. Special thanks to my good friend Vishnu for his intellectual support, and to my girlfriend Athira for her unwavering love and motivation throughout this journey. Their collective belief in me has been a driving force behind this dissertation.

References

Cho, J. H. & Hariharan, B. (2019), On the efficacy of knowledge distillation, *in* 'Proceedings of the IEEE/CVF international conference on computer vision', pp. 4794–4802.

Goodfellow, I. J., Shlens, J. & Szegedy, C. (2014), 'Explaining and harnessing adversarial examples', *arXiv preprint arXiv:1412.6572* .

Gou, J., Yu, B., Maybank, S. J. & Tao, D. (2021), 'Knowledge distillation: A survey', *International Journal of Computer Vision* **129**, 1789–1819.

He, K., Zhang, X., Ren, S. & Sun, J. (2016), Deep residual learning for image recognition, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 770–778.

Hinton, G., Vinyals, O. & Dean, J. (2015), 'Distilling the knowledge in a neural network', *arXiv preprint arXiv:1503.02531* .

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M. & Adam, H. (2017), 'Mobilenets: Efficient convolutional neural networks for mobile vision applications', *arXiv preprint arXiv:1704.04861* .

Krizhevsky, A. (n.d.), 'Cifar-10 dataset', https://www.cs.toronto.edu/~kriz/cifar.html. Accessed: 2023-08-24.

Krizhevsky, A., Hinton, G. et al. (2009), 'Learning multiple layers of features from tiny images'.

LeCun, Y., Bengio, Y. & Hinton, G. (2015), 'Deep learning', *nature* **521**(7553), 436–444.

Li, Z., Li, X., Yang, L., Zhao, B., Song, R., Luo, L., Li, J. & Yang, J. (2023), Curriculum temperature for knowledge distillation, *in* 'Proceedings of the AAAI Conference on Artificial Intelligence', Vol. 37, pp. 1504–1512.

Liu, J., Liu, B., Li, H. & Liu, Y. (2022), 'Meta knowledge distillation', *arXiv preprint arXiv:2202.07940* .

Papernot, N., McDaniel, P., Wu, X., Jha, S. & Swami, A. (2016), Distillation as a defense to adversarial perturbations against deep neural networks, *in* '2016 IEEE symposium on security and privacy (SP)', IEEE, pp. 582–597.

Polino, A., Pascanu, R. & Alistarh, D. (2018), 'Model compression via distillation and quantization', *arXiv preprint arXiv:1802.05668* .

Stanton, S., Izmailov, P., Kirichenko, P., Alemi, A. A. & Wilson, A. G. (2021), 'Does knowledge distillation really work?', *Advances in Neural Information Processing Systems* **34**, 6906–6919.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. & Fergus, R. (2013), 'Intriguing properties of neural networks', *arXiv preprint arXiv:1312.6199* .

Zagoruyko, S. & Komodakis, N. (2016), 'Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer', *arXiv preprint arXiv:1612.03928* .

Zhao, S., Wang, X. & Wei, X. (2023), 'Mitigating the accuracy-robustness trade-off via multi-teacher adversarial distillation', *arXiv preprint arXiv:2306.16170* .