Leveraging Stronger Teachers for Efficient Deep Learning Model Deployment in Resource-Constrained Settings

Lindani Dlamini
School of Computer Science and Applied Mathematics
University of the Witwatersrand
Johannesburg, South Africa
1712359@students.wits.ac.za

Richard Klein
School of Computer Science and Applied Mathematics
University of the Witwatersrand
Johannesburg, South Africa
richard.klein@wits.ac.za

***Abstract*—Recent advancements in deep neural network (DNN) compression have sought to address the computational demands of deploying these models in resource-constrained devices. In this research, we explore the nuanced realm of knowledge distillation, specifically focusing on the transfer of knowledge from robust vision transformers into convolutional neural networks. We employ state-of-the-art CNN models such as the ResNet variants as students, distilling knowledge into them from powerful vision transformers. We observe that the discrepancy in predictions between the student and a robust teacher can hinder the knowledge distillation process, necessitating a shift towards just preserving the essential relations between teacher and student predictions. Our focus is on leveraging correlation-based losses, explicitly capturing inter-class relations while extending the relational match to the intra-class level. Despite the inherent challenges of distilling knowledge from stronger teachers, our experiments reveal a 5% average gain in top-1 accuracy for all students on CIFAR-10 and a 10% increase on CIFAR-100, indicating the effectiveness of our chosen approach.**

***Index Terms*—deep learning, strong teacher, knowledge distillation**

I. Introduction

Deep neural networks (DNNs) have become an increasingly popular machine learning framework in recent years. Their impressive ability to learn from data and make predictions has made them quite effective at a variety of tasks such as object detection [20], image classification [19], and semantic segmentation [21]. However, the processing power and storage capacity required by these models is quite enormous, which makes it nearly infeasible to deploy them on consumer-grade or edge devices where we might need to perform real-time inference. Large state-of-the-art models are currently hosted on powerful company servers and are accessed by users via the internet. This is effective for simple tasks like language translation, but it becomes impractical when sensitive data, such as medical information, needs to be transferred to the models for processing. Therefore, it is necessary to develop techniques to create smaller, more efficient models that can achieve comparable accuracy to their larger counterparts and can be deployed in environments with limited resources [22].

Identify applicable funding agency here. If none, delete this.

One such technique that has been proposed to obtain these efficient models with similar accuracy of these larger models is knowledge distillation, which involves training a smaller model, known as the student, to learn to produce the same outputs as the larger teacher model, given the same inputs [1]. In the process of knowledge distillation, the effective formulation and transfer of knowledge into the student model play a pivotal role. One way to achieve this is by utilizing the probabilities generated by the teacher as ”soft targets” to train the student. Subsequently, the alignment of probability distributions between the teacher model and the student model is achieved through the utilization of the Kullback-Leibler (KL) divergence. This alignment helps ensure that the student model's predictions resemble those of the teacher model, even if they are not an exact match. Essentially, it helps the student model learn not just from the hard labels (actual targets) but also from the softer, more nuanced information that the teacher model provides. This allows the student model to capture the patterns and relationships present in the data more effectively, ultimately improving its performance and accuracy.

Besides this vanilla technique to match the teacher and student predictions, there exists a plethora of research and experiments [10]–[12] in response to this challenge of finding techniques for obtaining lightweight models. However, recent studies [13]–[15] have shown that the student struggles to learn the teacher's probabilities when there is a significant difference in size between them, which poses a challenge in effectively implementing knowledge distillation. To address this problem, [14] proposed to bring in a moderate-sized teaching assistant to help bridge the size gap. However, bringing in an additional model during training demands more compute resources. Hence, in this study, our focus lies on maintaining the teacher's preferences, specifically the predictions' relative ranks, rather than striving for their precise replication. This approach is adopted to enhance the efficiency of the knowledge distillation process. [16]. Our experiments on image classification using the CIFAR-10 and CIFAR-100 datasets [8] have shown that this approach can improve the accuracy of the student model.

II. Background

In this section, we begin by discussing deep neural networks and some of their most popular architectures. In particular, we shall focus on convolutional neural network architectures and vision transformers in image classification. We shall then move on to discussing knowledge distillation from a vision transformer into a convolutional neural network to provide an understanding of the subject.

A. Convolutional Neural Networks

Convolutional neural networks (CNNs) have become quite popular in image classification tasks more than the traditional fully connected neural networks because of their better performance with regard to processing data with a known grid-like structure such as an image. CNNs are still neural networks but they use convolution instead of traditional matrix multiplication in at least one of their layers [2]. A convolution is an operation on two functions, the input image and the kernel, to produce a feature map to extract the input image's features. The more convolutional layers we have, the more high-level features we extract to better classify the image. They are easier to train than traditional neural networks because of sparse connectivity and parameter sharing, which reduce operations carried out during training.

An inherent problem with deep neural networks is that they can be difficult to train and optimize given their millions of parameters. Reference [3] proposed a solution called deep residual learning, where the output of a layer is redefined as the sum of its input and the output of a residual function learned by the model as it trains. Their framework also addresses the degradation problem of deep neural networks by using feed-forward neural networks with "shortcut connections" to realize this formulation. Much empirical evidence is provided, proving the ease in optimization of such residual networks as well as the gains in accuracy achieved by increasing their depth. Image classification experiments were conducted on the ImageNet [23] and CIFAR-10 [8] datasets to evaluate the performance of their residual nets with different levels of depth. The Resnet is still quite the state-of-the-art (SOTA) architecture to researchers in the modern day and some proposed models built from it have achieved SOTA levels of accuracy [4].

B. The Vision Transformer (ViT)

While the above architectures were built on CNNs, the Vision Transformer (ViT), was built on the Transformer architecture proposed by [5] that is commonly used in natural language processing (NLP). Reference [6] extended or built upon the idea of using the Transformer architecture for image analysis. They applied the Transformer architecture directly to images by dividing an image into smaller patches and converting each patch into a linear embedding. These linear embeddings are then used as input to the Transformer model, allowing it to capture spatial relationships and dependencies within the image, similar to how it captures sequential information in natural language processing tasks. The authors evaluated the ability of ResNet and the Vision Transformer (ViT)to learn meaningful representations or features from raw data, pre-training these models on datasets of varying sizes and evaluating them on several benchmark tasks. They used several datasets, including CIFAR-10 and CIFAR-100 [8] to name a few, and found that even though ViT required less data and computation to train than other models, it was able to achieve high levels of performance across a majority of image recognition metrics.

C. Knowledge Distillation

In the pursuit of creating compact and efficient networks, two primary methods have emerged as the most effective. The first method involves directly reducing the size of a network through quantization and/or pruning of its parameters with the aim of reducing the computational complexity of a model without significantly impacting its performance [17]. The second is knowledge distillation, where we wish to train a compact model that is more suited for deployment and has the generalization capabilities of the larger model on new data. This can be achieved by transferring the knowledge of the larger model after it has been trained, into the smaller model. Initially, the teacher undergoes training with a specified dataset. The main idea is to then make the predictions of the student similar to those of the teacher. Given a batch size b and k classes, we can represent the teacher's logits as $\boldsymbol{Z}^{(\mathrm{t})} \in \mathbb{R}^{b\times k}$ and the student's logits as $\boldsymbol{Z}^{(\mathrm{s})} \in \mathbb{R}^{b\times k}$. It follows that we can express the usual KD loss [1] as

$$\mathcal{L}_{\mathrm{KD}} := \frac{\tau^2}{b}\sum_{i=1}^{b} \mathrm{KL}\left(\boldsymbol{Y}_{i,:}^{(\mathrm{t})}, \boldsymbol{Y}_{i,:}^{(\mathrm{s})}\right) \tag{1}$$

where KL symbolizes the Kullback-Leibler divergence, a metric quantifying the deviation of one probability distribution from another, the latter being the expected distribution. In this context, it quantifies the difference between the teacher's predicted probabilities and the student's predicted probabilities. The formula for the KL divergence uses the softmax of the teacher and student logits,

$$\boldsymbol{Y}_{i,:}^{(\mathrm{t})} = \mathrm{softmax}\left(\boldsymbol{Z}_{i,:}^{(\mathrm{t})}/\tau\right), \quad \boldsymbol{Y}_{i,:}^{(\mathrm{s})} = \mathrm{softmax}\left(\boldsymbol{Z}_{i,:}^{(\mathrm{s})}/\tau\right), \tag{2}$$

which are essentially the outputs of the models before the final activation function. The temperature parameter τ is used to control how 'soft' the logits are.

In the training process, both the classification loss (comparing the student's predictions with the ground-truth labels) and the KD loss are taken into account. This results in the training loss being a combination of the classification loss, denoted as $\mathcal{L}$cls , and the KD loss, denoted as $\mathcal{L}_{\mathrm{KD}}$., i.e.

$$\mathcal{L}_{\mathrm{tr}} = \alpha\mathcal{L}_{\mathrm{cls}} + \beta\mathcal{L}_{\mathrm{KD}} \tag{3}$$

where the relative importance of each loss is controlled by the factors α and β. The cross-entropy loss is typically used as the classification loss, and it measures the difference between the predicted class probabilities and the true distribution of classes.

The aim of this process is to enable the student model to capture the important patterns and relationships present in the data, leading to improved performance and accuracy. However, there exists key components that affect the efficacy of this process that need to be considered:

- Impact of Teacher and Student Model Strength: The strength of a model is determined by its ability to perform the task at hand effectively, often in comparison to other models. Several factors contribute to the strength of a teacher model:
  - Model Complexity: A strong model is often a complex model with a high representational capacity, i.e. it can capture intricate patterns and relationships within the data. This complexity can be achieved through various means, such as increasing the model size or by using more advanced architectural designs.
  - Training Strategies: Employing advanced training strategies, such as label smoothing, data augmentation, mix-up, or other regularization techniques, can enhance the generalization and robustness of the model. These strategies help the model learn from the data more effectively and improve its performance on the task.
  - Transferable Knowledge: A strong teacher model should possess comprehensive and transferable knowledge that can be effectively conveyed to the student throughout the process of knowledge distillation. This knowledge includes not only accurate predictions but also a nuanced understanding of the data that enables the model to make reliable and informed decisions.
- Limitations of Representational Capacity: Despite employing similar advanced training strategies, [16] showed from their experiments that the restricted representational capacity of a student hinders its ability to match the performance of a stronger teacher model. This leads to a widening performance gap between the two models, emphasizing the significance of architecture and model size in determining a model's learning capabilities.
- Challenges in KD with Stronger Training Strategies: The investigation by [16] revealed that employing stronger training strategies during KD not only amplifies the discrepancy between teacher and student models but also intensifies the misalignment between the knowledge distillation loss and the classification loss. This misalignment was observed to disturb the student model's training process, leading to a significant reduction in its learning capacity and performance.

III. Related Work

While a substantial body of research has concentrated on knowledge distillation within similar architectures, particularly from one convolutional neural network (CNN) to another, our focus diverges. Instead of confining the knowledge transfer to analogous models, we explore the intriguing realm of distilling knowledge from a vision transformer into a convolutional neural network. This departure seeks to narrow the performance gap between these distinct architectures. Some cross-architecture knowledge distillation methods, as exemplified in [7], leverage sophisticated techniques. Notably, they employ tools like a partially cross attention projector and a group-wise linear projector to seamlessly map the student CNN's feature space into the teacher Vision Transformer's transformer and feature spaces. It's worth noting the potency and thoroughness of this approach and their experimental evidence showcased a substantial increase in student accuracy. However, it's paramount to acknowledge that such an approach primarily emphasizes certain facets of knowledge distillation, such as the internal feature maps of the models. This awareness sets the stage for our specific exploration into novel aspects of knowledge distillation, as outlined next.

In tandem with our exploration of knowledge distillation, recent studies [13], [18], and the findings from [14] question the conventional wisdom surrounding the relationship between teacher and student model sizes. While it's widely acknowledged that a strong teacher model, often characterized by increased complexity and representational capacity, can impart valuable knowledge to a student, nuances in this relationship are surfacing. The work by [15] challenges the assumption that a larger teacher model inherently leads to improved student performance. These findings prompt a re-evaluation of the role model sizes play in knowledge distillation. Our investigation also considers the impact of training strategies on the efficacy of knowledge distillation. The works of [13] demonstrated that advanced training strategies, such as label smoothing, data augmentation, and regularization techniques, are integral to enhancing a model's generalization and robustness. However, caution is warranted. In their study, [15] reveal that stronger training strategies during knowledge distillation might exacerbate discrepancies between teacher and student models, potentially compromising the learning capacity and performance of the student. This underscores the delicate balance required when employing these strategies.

The core of knowledge distillation lies in aligning the probability distributions of teacher and student models. As discussed earlier, the KL divergence, encapsulated in the vanilla knowledge distillation loss, measures the disparity between predicted probabilities. Extending this discussion, [18] and the insights from [14] offer perspectives on refining this alignment process. Exploring techniques to better match distributions and overcome challenges identified in [15] becomes pivotal for the success of knowledge distillation. While sequential knowledge distillation has been proposed as a potential solution to reduce the disparity in capacities between the student and teacher, [13] and the findings from [15] caution about its effectiveness.

Building upon the insights from [14] and the limitations of sequential knowledge distillation demonstrated in [18], we can identify the need for a one-size-fits-all approach when tackling the challenges posed by stronger teachers in knowledge distillation. Instead of attempting to customize solutions for each category of stronger teachers, whether characterized by larger

model sizes or the utilization of sophisticated training strategies, this study promotes a universal and adaptable approach. The reason behind this lies in the difficulty of achieving an exact match between teacher and student outputs, especially when faced with a significant difference in performance. The traditional approach of minimizing loss using Kullback-Leibler (KL) divergence becomes increasingly difficult in these situations, given the considerable disparity between the teacher and student.

IV. Methodology

Building upon the foundation laid in the background and related work, we delve into our methodology, which revolves around distilling knowledge from a Vision Transformer (ViT) into a Convolutional Neural Network (CNN) to enhance the performance of the latter in image classification tasks.

A. Datasets

To evaluate the performance of our models, we used the CIFAR-10 and CIFAR-100 datasets. The CIFAR-10 dataset consists of 60,000 labeled images in 10 mutually exclusive classes, with a fixed size of 32x32x3 color channels. The CIFAR-100 dataset contains 600 images in each of 100 classes, with the same image size as CIFAR-10 [8]. The CIFAR-100 serves as a much more challenging task for the models. The images were re-scaled to a height and width of 224 before passing them into the models because the transformer teachers used in this research were pre-trained on those image dimensions as specified by [6].

B. Models

For the strong teacher networks, we used the $google/vit-base-patch16-224-in21k$ and the $google/vit-large-patch16-224-in21k$ from Hugging Face's pre-trained $transformers$ library [9]. Strong in the sense that they are relatively much larger than the student networks. For the students, we focus on the ResNet family models—ResNet18, ResNet50, and ResNet152. This intentional selection stems from the shared architecture among these models, differing only in size. This uniformity facilitates a nuanced understanding of how each student model absorbs knowledge distilled from their respective teachers, unveiling insights into the transferability of knowledge across varying model sizes. and were trained on a strong training strategy. In particular, they come pre-trained and we retrain them on our specific downstream task of image classification on the above-mentioned datasets, of which they attain impressive accuracy. Since we are also interested in observing how KD is affected as we scale up the size of the teacher, the difference in sizes of these two models will play a crucial role in that. For the student models, we chose the ResNet18, ResNet50 and ResNet152. This is to observe how the student models receive the knowledge from the teacher as they also increase in size. Detailed information of the models is presented in Table II.

C. Model Training and Associated Training Strategies

Before embarking on the knowledge distillation phase, we initiate our investigation with a common, robust training strategy applied uniformly across all models. Employing a challenging training regimen ensures that each model faces a comparable level of difficulty, setting the stage for subsequent knowledge distillation experiments. Data augmentations, including random brightness, contrast, saturation adjustments, hue variations, and flipping, inject diversity into the training process. Additionally, we utilize the AdamW optimizer to optimize the model's weights during this preliminary training phase. The hyperparameter values presented in Table I are the outcome of meticulous experimentation aimed at achieving optimal performance. Adjustments to parameters such as epochs, batch size, and learning rate were systematically explored to enhance the robustness and efficiency of our models. This iterative process is pivotal in fine-tuning our models for subsequent knowledge distillation analyses. Following the initial training, we evaluate the models' accuracy on both CIFAR-10 and CIFAR-100. The results, summarized in Table II, showcase a direct relationship between model accuracy and size, setting the groundwork for our subsequent knowledge distillation analyses.

D. Knowledge Distillation

In the context of training machine learning models, the prediction scores of a model represent how confident it is about different classes. When trying to match predictions between a teacher and a student model in this research, we're interested in a more flexible approach. Instead of focusing on exact numerical values during prediction, [16] proposed DIST, a method to capture the relationships or relative ranks of predictions made by the teacher. This means that during the inference phase, we're less concerned with the precise probability values and more interested in understanding the order or ranking of predicted classes by the teacher. This is particularly important for transferring knowledge effectively from the teacher to the student.

To formalize this relaxed match, a suitable metric is chosen that quantifies the difference between two prediction vectors. In our emphasis on relations over exact matches, it is crucial that a relaxed match does not necessitate precise equality between the prediction vectors. The metric should maintain the semantic information inherent in the prediction vectors and safeguarding the fidelity of inference results without distorting their relationships. In practical terms, this means the matching process is robust to adjustments in the overall confidence levels of the teacher.

Finally, to implement this relaxed match, the Pearson's distance,

$$d_{\mathrm{p}}(\boldsymbol{x}, \boldsymbol{y}) := 1 - r_{\mathrm{p}}(\boldsymbol{x}, \boldsymbol{y}) \tag{4}$$

where $r_{\mathrm{p}}(\boldsymbol{x}, \boldsymbol{y})$ is the Pearson correlation coefficient, is used to measure the correlation between two sets of predictions, is adopted as the metric. In doing so, the linear correlation

TABLE I
TRAINING STRATEGIES FOR IMAGE CLASSIFICATION

Dataset	Model	Epochs	Batch size	Learning rate
CIFAR-10	googlevit-base-patch16-224-in21k	7	16	5.0×10^{-6}
	google/vit-large-patch16-224-in21k	7	16	5.0×10^{-6}
	ResNet152	20	32	1.0×10^{-4}
	ResNet50	20	32	7.5×10^{-5}
	ResNet18	20	32	7.5×10^{-5}
CIFAR-100	googlevit-base-patch16-224-in21k	10	16	5.0×10^{-6}
	google/vit-large-patch16-224-in21k	10	16	5.0×10^{-6}
	ResNet152	25	32	1.0×10^{-4}
	ResNet50	25	32	7.5×10^{-5}
	ResNet18	25	32	7.5×10^{-5}

TABLE II
MODEL PERFORMANCE ON IMAGE CLASSIFICATION

Model	**Params (Million)**	**CIFAR-10**		**CIFAR-100**	
		Top1 Acc. (%)	**Top5 Acc. (%)**	**Top1 Acc. (%)**	**Top5 Acc. (%)**
google/vit-base-patch16-224-in21k	86.0	98.33	99.98	91.34	98.86
google/vit-large-patch16-224-in21k	307.0	98.35	100.00	92.02	98.99
ResNet18	12.9	82.25	98.91	50.79	75.70
ResNet50	22.4	83.12	99.29	54.37	81.41
ResNet152	59.0	84.17	99.40	55.54	82.96

between the teacher and student predictions is maximized, specifically focusing on preserving the inter-class relation.

It follows that the relation can then be defined as a correlation. In this context, the inter-relation loss

$$\mathcal{L}_{\text{inter}} := \frac{1}{b}\sum_{i=1}^{b} d_{\text{p}}\left(\boldsymbol{Y}^{(\text{s})}_{i,:}, \boldsymbol{Y}^{(\text{t})}_{i,:}\right) \tag{5}$$

is established for every pair of prediction vectors $\boldsymbol{Y}^{(\text{s})}_{i,:}$ and $\boldsymbol{Y}^{(\text{t})}_{i,:}$, representing a crucial aspect of preserving relations.

In addition to considering the relationship between different classes for each instance, [16] also find value in understanding the prediction scores of multiple instances within the same class. Even within the same class, the variability in how similar instances are to that class is informative. Intra-class relation focuses on the correlation between prediction scores for the same class across multiple instances as can be seen in Fig 1.

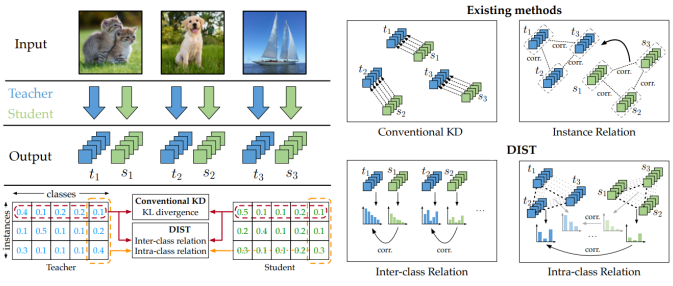

Fig. 1. Difference between DIST and Conventional KD. Adapted from [16]

If we represent the prediction scores in matrices, with each row corresponding to the predictions for one instance, the inter-class relation can be conceptualized as an effort to maximize the correlation row-wise. In contrast, the intra-class relation aims to maximize the correlation column-wise, i.e. it looks at how predictions for the same class are related across different instances. Hence this intra-relation is also distilled into the student for better performance:

$$\mathcal{L}_{\text{intra}} := \frac{1}{k}\sum_{j=1}^{k} d_{\text{p}}\left(\boldsymbol{Y}^{(\text{s})}_{:,j}, \boldsymbol{Y}^{(\text{t})}_{:,j}\right) \tag{6}$$

It follows that the total training loss $\mathcal{L}_{\text{tr}}$ is a combination of the original classification, inter-class and intra-class losses, i.e.

$$\mathcal{L}_{\text{tr}} = \alpha\mathcal{L}_{\text{cls}} + \beta\mathcal{L}_{\text{inter}} + \gamma\mathcal{L}_{\text{intra}} \tag{7}$$

as formulated by [16], and the factors α, β, and γ control the importance of each component. This approach, using both inter-class and intra-class relations, allows the student model to adaptively match the teacher's output, significantly improving the distillation performance.

V. EXPERIMENTS

A. Experimental Settings

1) Knowledge Distillation Training Strategies: For our knowledge distillation experiments, we adopt a streamlined training strategy, emphasizing simplicity and efficiency. The goal is to evaluate the efficacy of knowledge transfer with minimal complexity. The data augmentation during knowledge distillation is less intricate, comprising random cropping and flipping. This choice aligns with the rationale that a more straightforward training strategy can effectively distill knowledge while reducing computational overhead.

TABLE III
TRAINING STRATEGIES FOR KNOWLEDGE DISTILLATION

Dataset	Epochs	Batch size	Learning rate	α	β	γ	τ
CIFAR-10	20	16	7.5×10^{-5}	1	2	2	2
CIFAR-100	20	16	7.5×10^{-5}	1	2	2	4

TABLE IV
STUDENT MODEL PERFORMANCE POST KNOWLEDGE DISTILLATION

Dataset	Teacher	Student	Teacher acc.		Student acc.		Student acc.	
			Top1(%)	Top5 (%)	Top1 (%)	Top5 (%)	Top1 (%)	Top5 (%)
CIFAR-10	google/vit-base-patch16-224-in21k	ResNet18	98.33	99.98	82.25	98.91	87.26	99.15
		ResNet50	–	–	83.12	99.29	88.23	99.36
		ResNet152	–	–	84.17	99.40	88.86	99.37
CIFAR-10	google/vit-large-patch16-224-in21k	ResNet18	98.35	100.00	82.25	98.91	87.56	99.20
		ResNet50	–	–	83.12	99.29	88.84	99.41
		ResNet152	–	–	84.17	99.40	89.13	99.59
CIFAR-100	google/vit-base-patch16-224-in21k	ResNet18	91.34	98.86	50.79	75.70	60.46	85.55
		ResNet50	–	–	54.37	81.41	64.72	86.79
		ResNet152	–	–	55.54	83.96	65.32	88.84
CIFAR-100	google/vit-large-patch16-224-in21k	ResNet18	92.02	98.99	50.79	75.70	61.31	85.89
		ResNet50	–	–	54.37	81.41	65.39	87.54
		ResNet152	–	–	55.54	83.96	67.32	89.27

2) Hyperparameters: The knowledge distillation process involves careful selection of hyperparameters. In Table III, we present the specific values used for epochs, batch size, learning rate, α, β, γ, and τ. We opt for a higher temperature, $\tau = 4$, on CIFAR-100. This choice is motivated by the tendency to encounter overfitting and sharp learned probabilistic distributions on CIFAR-100 [16].

Notably, α and β control the importance of classification loss and knowledge distillation loss, while γ modulates the effect of label smoothing. The temperature parameter τ adjusts the softness of the predicted distributions. Our choice of hyperparameters aims to strike a balance between effective knowledge transfer and model generalization.

B. Results

In Table IV, we present the results of knowledge distillation experiments where ResNet models (ResNet18, ResNet50, and ResNet152) were used as students. The aim was to distill knowledge from Google's Vision Transformer models with varying sizes. We observed an average increase of 5% in top-1 accuracy across all students on CIFAR-10. This increase was similar under both teachers. For CIFAR-100, we observed an average increase of 10% in top-1 accuracy across all students when learning from $google/vit - base - patch16 - 224 - in21k$, and a similar increase of about 11% when learning from $google/vit - large - patch16 - 224 - in21k$.

The results highlight intriguing trends. Despite larger teachers presenting challenges for vanilla knowledge distillation, DIST, showcases an upward trend in student performance as teacher size increases. There is a slightly greater increase in accuracy across all students when comparing the gains from the larger $google/vit-large-patch16-224-in21k$ teacher. This improvement is particularly noteworthy, emphasizing the effectiveness of DIST in mitigating the challenges posed by larger teacher models. Overall, the experiments validate the robustness and adaptability of our knowledge distillation strategy, shedding light on its potential for facilitating knowledge transfer across varying model sizes and training complexities. We also note that [16] conducted rigorous experiments in their study to demonstrate that DIST outperforms conventional KD methods, hence it was directly applied in this work.

VI. CONCLUSION

In this study, we delved into the intricate landscape of knowledge distillation, with a specific focus on transferring knowledge from powerful vision transformers to convolutional neural network students. Our exploration aimed to bridge the performance gap between these disparate architectures. Leveraging state-of-the-art CNN models, including ResNet18, ResNet50, and ResNet152, as students, we distilled knowledge from robust vision transformers, such as $google/vit - base - patch16 - 224 - in21k$ and $google/vit - large - patch16 - 224 - in21k$.

Our chosen approach, termed DIST [16], introduced a relaxed match with relations, emphasizing the preservation of inter-class relations rather than exact probabilistic values during knowledge transfer. We employed Pearson's distance as a metric to measure the correlation between teacher and student predictions, resulting in an inter-relation loss. This nuanced approach demonstrated its effectiveness, particularly when faced with challenges posed by larger teacher models.

The experiments, conducted on CIFAR-10 and CIFAR-100 datasets [8], revealed promising trends. Despite the inherent difficulties associated with distilling knowledge from stronger teachers, DIST showcased an encouraging improvement in student performance as the size of the teacher increased. This suggests that it is robust and adaptable, providing insights into effective knowledge transfer across varying model sizes and training complexities. The importance of preserving inter-class relations, as demonstrated by DIST, opens avenues for

further exploration and refinement of knowledge distillation techniques.

VII. Future Work

The exploration of knowledge distillation's transferability across diverse neural network architectures is paramount to understanding its versatility. Extending the application of the proposed approach beyond ResNet, for instance, to architectures like EfficientNet or MobileNet to name a few, presents an opportunity to glean insights into the generalizability of the knowledge distillation strategy. Additionally, investigating the impact of employing even larger teachers, such as $google/vit-huge-patch16-224-in21k$, would provide valuable observations on the dynamics of knowledge transfer across a broader spectrum of model sizes.

References

[1] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *ArXiv*, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:7200347.

[2] I. Goodfellow, Y. Bengio, and A. Courville, "Deep learning: Adaptive computation and machine learning," *MIT Press*, 2017.

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770-778. doi: 10.1109/CVPR.2016.90.

[4] Papers With Code, "Deep Residual Learning for Image Recognition," 2016. [Online]. Available: *https://paperswithcode.com/paper/deep-residual-learning-for-image-recognition*. Accessed: 2023-04-19.

[5] A. Vaswani et al., "Attention is All You Need," 2017. [Online]. Available: *https://arxiv.org/pdf/1706.03762.pdf*. Accessed: 2023-04-25.

[6] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *ArXiv*, 2020. [Online]. Available: *https://arxiv.org/abs/2010.11929*.

[7] Y. Liu et al., "Cross-Architecture Knowledge Distillation," 2022. [Online]. Available: *https://arxiv.org/abs/2207.05273*.

[8] A. Krizhevsky, V. Nair, and G. Hinton, "The CIFAR-10 dataset," 2014. [Online]. Available: *https://www.cs.toronto.edu/ kriz/cifar.html*.

[9] T. Wolf et al., "Transformers: State-of-the-Art Natural Language Processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Oct. 2020, pp. 38–45. [Online]. Available: *https://www.aclweb.org/anthology/2020.emnlp-demos.6*.

[10] T. Nguyen-Duc, T. Le, H. Zhao, J. Cai, and D. Phung, "Adversarial Local Distribution Regularization for Knowledge Distillation," in *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA, 2023, pp. 4670-4679. doi: 10.1109/WACV56688.2023.00466.

[11] Z. Li, B. Yang, P. Yin, Y. Qi, and J. Xin, "Feature Affinity Assisted Knowledge Distillation and Quantization of Deep Neural Networks on Label-Free Data," *IEEE Access*, vol. 11, pp. 78042-78051, 2023. [Online]. Available: *https://api.semanticscholar.org/CorpusID:257079264*.

[12] I. Mishra, S. V. Krishna, and D. Mishra, "Distilling Calibrated Student from an Uncalibrated Teacher," *ArXiv*, vol. abs/2302.11472, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:257079238.

[13] J. H. Cho and B. Hariharan, "On the Efficacy of Knowledge Distillation," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), 2019, pp. 4793-4801. doi: 10.1109/ICCV.2019.00489.

[14] S. I. Mirzadeh, M. Farajtabar, A. Li, and H. Ghasemzadeh, "Improved Knowledge Distillation via Teacher Assistant: Bridging the Gap Between Student and Teacher," *ArXiv*, vol. abs/1902.03393, 2019. [Online]. Available: *https://api.semanticscholar.org/CorpusID:60440652*.

[15] W. Son, J. Na, J. Choi, and W. Hwang, "Densely Guided Knowledge Distillation using Multiple Teacher Assistants," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, 2021, pp. 9375-9384. doi: 10.1109/ICCV48922.2021.00926.

[16] T. Huang, S. You, F. Wang, C. Qian, and C. Xu, "Knowledge Distillation from A Stronger Teacher," *ArXiv*, vol. abs/2205.10536, 2022. [Online]. Available: *https://api.semanticscholar.org/CorpusID:248986690*.

[17] K. Paupamah, S. James, and R. Klein, "Quantisation and Pruning for Neural Network Compression and Regularisation," in *2020 International SAUPEC/RobMech/PRASA Conference*, Cape Town, South Africa, 2020, pp. 1-6. doi: 10.1109/SAUPEC/RobMech/PRASA48453.2020.9041096.

[18] Y. Xiong, W. Zhai, X. Xu, J. Wang, Z. Zhu, C. Ji, J. Cao, "Ability-aware knowledge distillation for resource-constrained embedded devices," *Journal of Systems Architecture*, vol. 141, 2023, article number 102912, ISSN 1383-7621, *https://doi.org/10.1016/j.sysarc.2023.102912*.

[19] T. Huang, S. You, B. Zhang, Y. Du, F. Wang, C. Qian, and C. Xu, "Dyrep: Bootstrapping training with dynamic re-parameterization," *arXiv preprint arXiv:2203.12868*, 2022.

[20] Z. Cai and N. Vasconcelos, "Cascade R-CNN: High Quality Object Detection and Instance Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2019.

[21] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid Scene Parsing Network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2881–2890.

[22] Neptune.ai, "Building and Deploying Computer Vision Models," 2023. [Online]. Available: *https://neptune.ai/blog/building-and-deploying-cv-models*.

[23] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-scale Hierarchical Image Database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, 2009, pp. 248-255. doi: 10.1109/CVPR.2009.5206848.