Transferring Vision: Teaching CNNs to See with ViT Wisdom

Kris Jeong
Stanford University
Department of Computer Science
kjeong@stanford.edu

Pauline Arnoud
Stanford University
Department of Computer Science
parnoud@stanford.edu

Abstract

This paper explores the use of Vision Transformers (ViTs) as teacher models to enhance the training efficiency and performance of Convolutional Neural Networks (CNNs) through a knowledge distillation process. We replicate the distillation strategy proposed by Touvron et al., where a CNN teacher model is used to distill knowledge into a ViT student model. Inspired by their approach, we reverse the roles by employing a pretrained ViT as the teacher and a ResNet-50 as the student model. Our methodology adapts their original distillation loss function, optimizing the student CNN to learn from both the ground truth labels and the teacher ViT's predictions. This approach aims to combine the computational efficiency of CNNs with the advanced learning capabilities of ViTs. Experimental results on the CIFAR-100 dataset indicate that the distilled ResNet-50 demonstrates a steady increase in accuracy, suggesting the potential to surpass the baseline ResNet-50 with extended training. This study contributes to the understanding of cross-architecture knowledge distillation and offers insights for future research in efficient model training.

1. Introduction

In recent years, transformers have demonstrated remarkable success in computer vision tasks, significantly outperforming conventional models. However, their extensive resource requirements and prolonged training times pose considerable challenges. Vision Transformers (ViTs), while highly effective, demand substantial computational power and time, which can be prohibitive for many researchers and practitioners [1].

A notable attempt to address these challenges is the teacher-student model proposed by Touvron et al. [2], which leverages a more efficient and well-performing model, such as RegNetY-16GF [3], to facilitate the training of ViTs. This approach successfully achieved comparable accuracy with significantly reduced training time, specifically 73 hours. Despite this improvement, the training duration remains a barrier for those with limited computational resources, particularly when working with smaller, custom datasets.

To further enhance training efficiency, we investigate whether the same strategy can be applied to Convolutional Neural Networks (CNNs), which are generally more efficient to train compared to ViTs due to their lower computational complexity and inherent inductive biases [4]. Specifically, we explore the possibility of utilizing a pretrained ViT as a teacher model to guide the training of a CNN. Our hypothesis is that this additional guidance from the teacher model will lead to improved performance or faster training of the CNN.

In this study, we employ a distillation training algorithm that adapts and modifies the approach proposed by Touvron et al. to suit CNNs. We train a CNN on a single 8-GPU node over 12 hours, leveraging the timm library [5] to implement our original contributions. The key contributions of this paper are as follows:

- We introduce an original CNN-specific distillation procedure inspired by the transformer-based distillation method of Touvron et al., effectively reversing the roles of the teacher and student models. Our model learns from both the ground truth labels and the teacher's predictions, optimizing a weighted sum of these inputs.
- Experimental results demonstrate that while our distilled ResNet-50 did not outperform the baseline ResNet-50 within the given epoch limit, the steady improvement suggests potential for better performance with extended training. We compare and contrast these results with Touvron et al.'s findings, generating insights about the structural differences between CNNs and ViTs and their use cases.

This study aims to explore the feasibility of integrating the advantageous properties of transformers into CNNs and to derive insights regarding cross-architectural knowledge transfer.

2. Related Works

Since their introduction in 2017 by Vaswani et al. [6], transformers have established themselves as the reference model for natural language processing. Dosovitskiy et al. extended this paradigm to computer vision with the introduction of Vision Transformers (ViTs) [1]. ViTs remove the need for convolutions by segmenting images into patches and processing these patches as sequences through self-attention mechanisms. This architecture excels in capturing long-range dependencies, surpassing state-of-the-art Convolutional Neural Networks (CNNs) in both accuracy and efficiency on various benchmarks. However, a key limitation of this work was that training of these models was resource-intensive, relying on the extensive JFT-300M dataset [7] (which consists of 300 million images) and did not generalize well with limited resources.

To address these limitations, Touvron et al. [2] leveraged Hinton et al.'s work on Knowledge Distillation [8] to introduce a specialized distillation procedure for ViTs. Their approach integrates an additional class token, termed the "distillation token," enabling the student ViT to learn from both the teacher CNN's predictions and the ground truth labels from ImageNet. Utilizing RegNetY-16GF as the teacher model, their data-efficient image transformer (DeiT) achieved comparable accuracy on ImageNet to the original ViT without any external pretraining data and maintained competitive performance across various downstream tasks, including CIFAR-10, CIFAR-100, Oxford-102 Flowers, Stanford Cars, and iNaturalist-18/19. [9, 10, 11, 12].

Despite the advancements of ViTs, CNNs have retained their relevance due to their computational efficiency and competitive performance on smaller datasets. CNNs, introduced with AlexNet in 2012 [13], have been the standard for image classification tasks due to their lower computational complexity, smaller parameter space, and inherent inductive biases such as translation invariance, which facilitate more efficient training compared to ViTs.

The integration of transformer-like attention mechanisms into CNNs has been a subject of ongoing research. Several studies have proposed architectures that leverage attention mechanisms within CNNs [14, 15], some even designing directly transplanting transformer components into CNNs [16, 17]. Building upon these efforts, our work proposes a novel approach inspired by Touvron et al.'s distillation model. Specifically, we employ a pretrained ViT as a teacher model to guide the training of a student CNN, aiming to achieve enhanced performance with more resource-efficient training.

3. Datasets

For our project, we used the CIFAR-100 dataset from the Canadian Institute for Advanced Research (CIFAR), introduced by Krizhevsky and Hinton in their 2009 technical report, "Learning multiple layers of features from tiny images" [9]. The CIFAR-100 dataset is widely used for benchmarking image classification algorithms, and consists of 60,000 32x32 color images across 100 classes, with 600 images per class. As we are investigating resource-efficient training, we wanted a dataset that was smaller than ImageNet or JFT-300M. We split this dataset into 50,000 training images and 10,000 test images. All the images are of a fixed resolution of 32x32 pixels ensuring uniformity across the dataset and facilitating efficient training and evaluation of our models. For each of our models, we used PyTorch's torchvision.datasets module to load the CIFAR-100 dataset and create data loaders for the training and validation sets.

To augment the training data and improve the robustness of our models, we applied several preprocessing steps and transformations. For the training set, we performed random cropping with a padding of 4 pixels, followed by random horizontal flipping. These augmentations help in simulating variations in the dataset, thus enhancing the model's generalization capabilities. Additionally, the images were normalized using the mean and standard deviation values of the CIFAR-100 training set: [0.5071, 0.4865, 0.4409] for the mean and [0.2673, 0.2564, 0.2762] for the standard deviation. For the validation set, we applied only normalization using the same mean and standard deviation values.

In the original DeiT (Data-efficient Image Transformers) model [2], the RegNetY-160 CNN teacher is initialized with weights pretrained on the ImageNet dataset [18], which consists of 1.2 million natural images across 1,000 classes. To adapt the RegNetY-160 teacher model to the CIFAR-100 dataset, we modified the final linear layer to have 100 output classes instead of the original 1,000 classes for ImageNet. Similarly, in our new proposed DeiT architecture, we initialized our ViT teacher with weights pretrained on the ImageNet dataset. This pre-training on the large-scale ImageNet dataset allows the ViT to learn rich visual representations that we wanted to take advantage of so they might be transferred and distilled into the CIFAR-100 domain. We therefore kept the original ViT model architecture used in the paper and, to accommodate the CIFAR-100 dataset, bilinearly interpolated the 32x32 images to 224x224 to match the expected input resolution of the pretrained ViT student model.

4. Methods

Our methodology begins with a simplified recreation of Touvron et al.'s architecture to establish a baseline on CIFAR-100, verifying the reproducibility of their results on our machines. We then implement our original approach. We visualize these architectures in Figures 2 and 3 respectively.

4.1. Recreating Touvron et al.'s Implementation

The Vision Transformer (ViT) architecture, as proposed by Dosovitskiy et al. [1], processes images by segmenting them into non-overlapping patches. For an image of size $H \times W$ with patch size $P \times P$, the image is divided into $\frac{H \times W}{P^2}$ patches. Each patch is flattened into a vector and linearly projected into a D-dimensional embedding space. To retain positional information, learnable position embeddings are added to these patch embeddings. The sequence of patch embeddings is then processed by a standard Transformer encoder, which comprises alternating layers of multi-head self-attention and feed-forward networks (FFNs) with residual connections and normalization.

In the ViT architecture, the class token is a trainable vector, appended to the patch tokens before the first layer. This token traverses the transformer layers and is then projected with a linear layer to predict the class. This class token, inherited from NLP models, departs from the typical pooling layers used in computer vision to predict the class. The transformer processes batches of $(N + 1)$ tokens of dimension D, of which only the class vector is used to predict the output. This architecture forces the self-attention to spread information between the patch tokens and the class token: at training time, the supervision signal comes only from the class embedding, while the patch tokens are the model's only variable input.

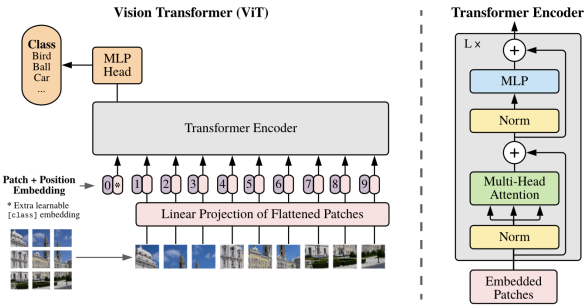

Figure 1. Architecture of the Vision Transformer (ViT) as proposed by Dosovitskiy et al. [1].

Touvron et al. extended this model by augmenting the original ViT with an additional class token termed the "distillation token." The class token learns to replicate the ground truth labels, while the distillation token learns to replicate the teacher model's predictions. During training, these two tokens' cross-entropy losses are combined in a weighted sum, and during inference, the average of these predictions is returned. By incorporating a token specifically focused on mimicking the teacher's outputs, the student transformer can more effectively distill the teacher's knowledge into its own parameters.

Touvron et al. explored two methods of distillation: soft distillation and hard distillation. Soft distillation minimizes the Kullback-Leibler divergence between the softmax of the teacher and the softmax of the student model. Hard distillation takes the teacher model's class predictions as the ground truth labels for the distillation token. We implemented the latter, which Touvron et al. reported consistently outperformed the former. We therefore used the following loss function:

$$L_{\text{global}}^{\text{hardDistill}} = \frac{1}{2} L_{CE}(\psi(Z_s), y) + \frac{1}{2} L_{CE}(\psi(Z_s), y_t).$$

where Z_s are the logits from the student model, $y_t = \arg\max_c Z_t(c)$ is the prediction of the teacher model, y is the ground truth labels, and ψ is the softmax function.

As for the selection of teacher models, Touvron et al. experimented with several CNNs and ViTs and concluded that the RegNet Y-16GF yields the best performance. Their implementation imports this model from the timm library and loads in their pretrained weights. In our recreation, we replaced the classifier head to match the CIFAR-100 dataset.

For the student model, Touvron et al. experimented with ViTs with different parameter combinations. In our recreation, we chose to replicate the DeiT-base model (deit-base-distilled-patch16-224), as it had the best accuracy compared to models with a smaller parameter space. The architecture is identical to the original paper except for the classifier head, which we modified to output the correct number of classes for CIFAR-100.

4.2. Developing Our Architecture

Subsequently, we developed our original architecture using a ViT teacher and a CNN student. We used the same pipeline we built to recreate Touvron et al.'s work, with the following changes.

First, for the teacher model, we selected the deit-base-patch16-224 model, which Touvron et al. trained on ImageNet and provided the trained weights for. We adjusted the classifier head to output the correct number of classes for CIFAR-100, resizing input images to 224x224.

For the student model, we chose ResNet-50 and trained it from scratch. Given our project's aim to investigate performance improvements with less training, using a more complex CNN would likely obscure the impact of the teacher. Thus, a simpler model like ResNet-50 allows clearer observation of the teacher's influence.

Our 'SwitchedDistillationLoss' class combines the Cross Entropy Loss between the student's predictions and the ground truth labels with the Cross Entropy Loss between the student's and teacher's predictions. The loss used for backpropagation is a weighted sum of these two components. A key design decision we made is to have the student model make a single prediction as opposed to two predictions (a "distillation map" and a regular activation map),

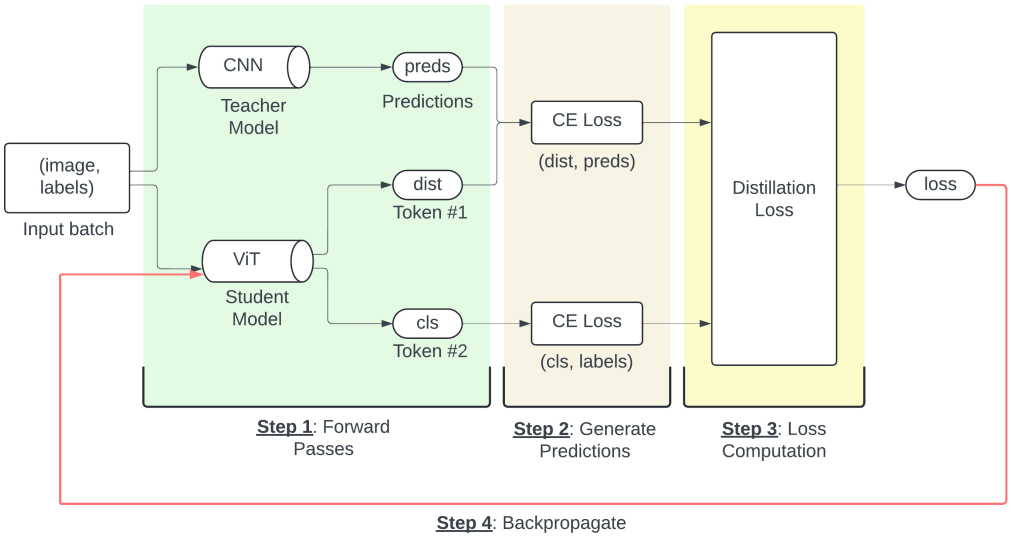

Figure 2. Diagram of Touvron et al.'s distillation model.

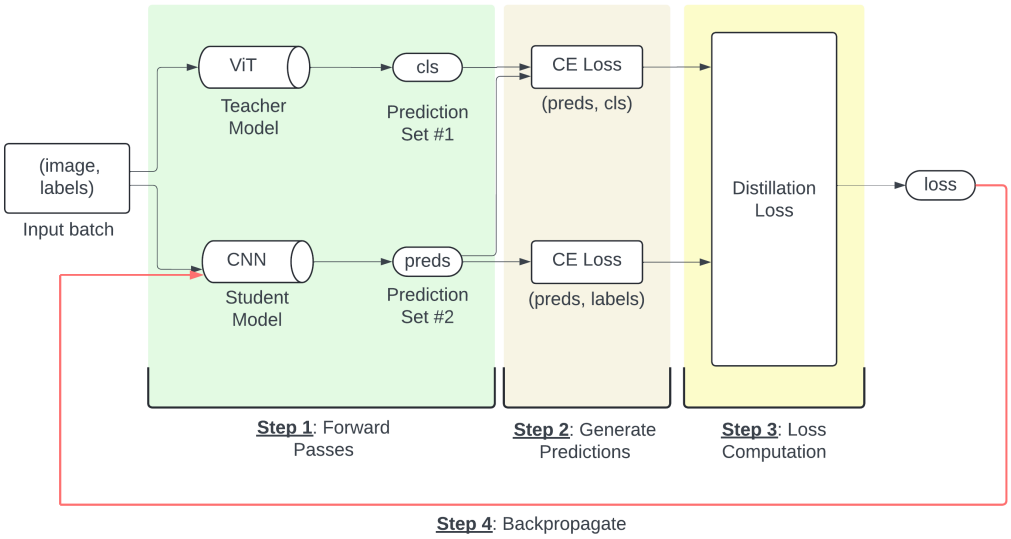

Figure 3. Diagram of our proposed switched teacher-student model.

which would be more aligned with Touvron et al.'s approach. This is because our student model is a CNN, which makes the implementation of a separate distillation token less straightforward and potentially less effective. CNNs are typically more effective when the complexity of the model is reduced to match their inductive biases (e.g., local receptive fields). Introducing a distillation token could disrupt this balance, making the training less efficient. By using a single output, we ensure that the CNN's architecture remains optimized for its strengths, while still benefiting from the teacher model's distilled knowledge.

By synthesizing the efficiency of CNNs with the advanced learning capabilities of ViTs through distillation, our methodology aims to enhance performance while reducing resource consumption.

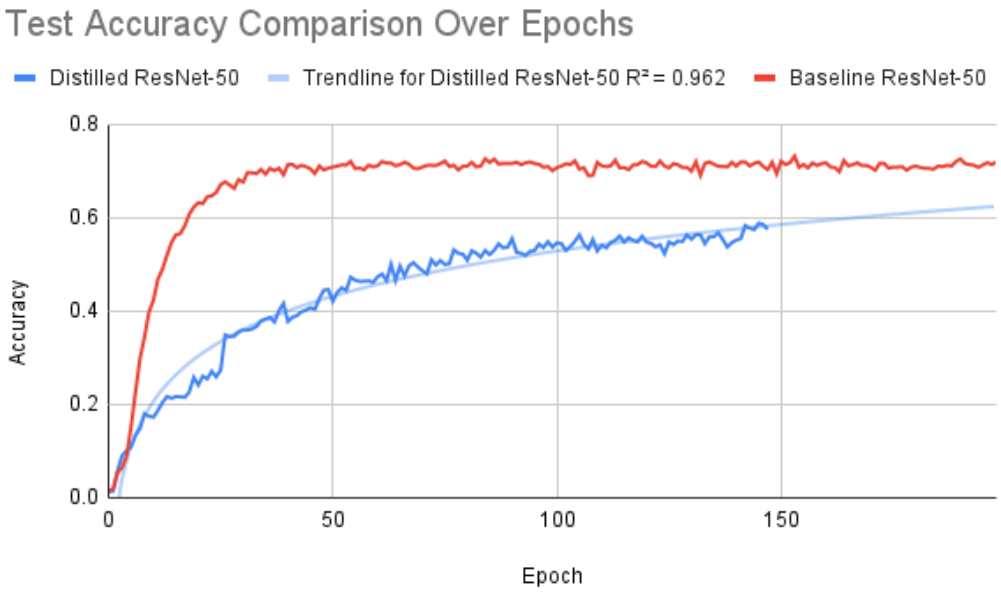

Figure 4. Test accuracy of baseline ResNet-50 and distilled ResNet-50.

5. Experiments & Results

5.1. Experimental Setup

In our experiment, we sought to validate the effectiveness of the distillation process by comparing the performance of a distilled ResNet-50 against a vanilla ResNet-50 baseline. To ensure the reliability and reproducibility of our results, we referenced Wightman et al.'s comprehensive survey on training strategies for ResNet models [19], adopting their hyperparameters to train our baseline ResNet-50. This approach helped us establish a robust baseline, ensuring that any observed differences in performance could be attributed to the distillation process rather than anomalies in the training setup.

Wightman et al. reported a Top-1 accuracy of 75.3% for a ResNet-50. In our experiments on CIFAR-100, our baseline ResNet-50 achieved a Top-1 accuracy of approximately 72%. We then applied the same hyperparameters to train the distilled ResNet-50, maintaining consistency across experiments. This consistency in experimental setup ensures that any differences in results can be attributed to the only variable we modified: the incorporation of the distillation loss, which measures how closely our student model mimics the outputs of the teacher ViT.

Our primary metric for evaluation was Top-1 accuracy, which measures the percentage of test examples where the model's highest confidence prediction matches the ground truth label. We picked this metric because it directly reflects the model's predictive power and offers a strict and clear measure of correctness, avoiding the leniency of metrics like top-5 accuracy. Top-1 accuracy is also a standard benchmark in image classification, and the metric used by Touvron et al. [2]. Thus, by focusing on top-1 accuracy, we ensured that we were making meaningful comparisons with existing research and directly assessing the impact of the distillation process. The hyperparameters used in our experiments are detailed in Table 5.1.

Parameter	Value
Model	ResNet-50
Batch Size	128
Epochs	2000
Seed	1
Criterion	`nn.CrossEntropyLoss`
Num Workers	1
Input Size	32
Num Classes	100
Learning Rate (LR)	0.01
Optimizer	SGD
Momentum	0.9
Weight Decay	0.0005
Scheduler	Cosine
T_max	N/A
Pretrained	FALSE

Table 1. Hyperparameters

Our distilled ResNet-50 model was trained for approximately 19 hours, which allowed for roughly 150 epochs. The baseline model ran for 200 epochs over 5 hours 42

minutes. The discrepancy in epoch duration is likely attributable to the additional computational overhead introduced by the distillation process, especially as a transformer's forward pass generally takes longer than a CNN's due to its complex architecture and larger parameter space.

5.2. Results and Analysis

We evaluated the models at each epoch, plotting the test accuracy against the number of epochs. To analyze the trends in accuracy improvement, we fitted logarithmic trendlines to the accuracy data. The baseline ResNet-50 had an R^2 score of 0.662, suggesting that the baseline model's accuracy was not following a logarithmic trend but rather plateaued around the 72% mark. In contrast, the distilled ResNet-50's accuracy trendline exhibited a strong fit, with an R^2 score of 0.962. The logarithmic trendline for the distilled ResNet-50's accuracy is given by the equation:

$$y = -0.112 + 0.139 \ln(x)$$

where y represents the test accuracy and x the number of epochs. Extrapolating this trendline, we believe the accuracy of the distilled ResNet-50 will continue to increase and eventually overtake the vanilla ResNet-50 around epoch 400. The performance comparison between the two models is illustrated in Figure 4. The baseline model's performance plateaued early, while the distilled ResNet-50 showed a steady increase in accuracy. Although the distilled model did not surpass the baseline within the 150 epochs, the trend suggests potential for continued improvement and possibly overtaking the baseline with extended training.

Using Figure 4, we can analyze how Vision Transformers (ViTs) impact the training of ResNet-50 in light of the structural aspects of ViTs and CNNs. Several noteworthy insights emerge from this experiment, especially when comparing our results against those of Touvron et al. [2].

First, the epochs for the distilled ResNet-50 took significantly longer than the baseline ResNet-50, likely due to the additional computations involved in distillation and the complexity of the ViT teacher model. The student model must not only compute its own predictions but also compare them against the teacher model's outputs. This comparison involves additional forward passes through the teacher model, which, being a ViT, further contributes to the increased computational time. The transformer's self-attention mechanism involves quadratic complexity in relation to the sequence length (i.e., the number of patches) [6]. Touvron et al. did not note an increase in training time compared to baseline ViTs because the time required for the additional forward pass through the CNN was negligible compared to a normal forward pass through a transformer. Our setup, involving a ViT as a teacher, introduces computational overhead that may not align well with the efficient training dynamics typically associated with CNNs.

Second, we notice that the distilled ResNet-50 model, for the number of epochs we were able to run, underperformed compared to the baseline; however, the distilled ResNet-50 shows steady logarithmic improvement with an R^2 value of 0.96, indicating consistent but slow learning, whereas the baseline ResNet-50 achieves rapid initial gains and plateaus around 72%. The slow but steady improvement in the distilled model aligns with Touvron et al.'s findings that distillation can progressively transfer knowledge, albeit requiring longer training durations to match or exceed the teacher's performance. This suggests that the baseline ResNet-50 is possibly overfitting on the training set by leveraging its inductive biases optimized for local feature learning [4]. The baseline model quickly learns the dominant features but struggles to generalize beyond them, highlighting the limitations of CNNs on this dataset without additional regularization. This tendency to overfit on local dependencies is mitigated in the distilled model due to the global attention the teacher ViT is able to impart. Given our trendline and Touvron et al.'s findings that distilled ViT models achieve high accuracy, often surpassing their CNN counterparts, we can optimistically predict that, with extended training, our distilled CNN could eventually surpass the baseline.

Several other factors inherent to the architecture of transformers and CNNs may contribute to the suboptimal performance of our distilled ResNet-50 for the initial 150 epochs. Firstly, using the same hyperparameters for both models ensured consistency, but this might not be optimal for the distilled model. Future work could explore hyperparameter tuning specifically for the distilled model, as the optimal settings might differ from those of the baseline due to the added complexity of the distillation process. Additionally, CIFAR-100 images (32x32 pixels) may not fully exploit the global attention mechanism of ViTs, which are designed to capture long-range dependencies [1]. Touvron et al. used larger datasets like ImageNet, where the global attention mechanism of ViTs can be fully utilized. The mismatch between the ViT's strengths and the CIFAR-100 dataset's characteristics could explain the suboptimal performance observed in our distilled ResNet-50. Future investigations should use datasets with larger images featuring more complex scenes to determine if the global dependencies leveraged by ViTs can outperform a basic CNN.

6. Conclusion

In this paper, we investigated the feasibility of improving the training efficiency and performance of Convolutional Neural Networks (CNNs) using a Vision Transformer (ViT) as a teacher model. Our approach was inspired by the successful distillation strategy proposed by Touvron et al., which utilizes a distillation token to transfer knowledge from a CNN teacher to a ViT student [2]. We reversed this

paradigm by employing a pretrained ViT as the teacher and a ResNet-50 as the student. While our distilled ResNet-50 did not outperform the baseline within the given epoch limit, the steady improvement suggests potential for better performance with extended training. The increased computational overhead and the specific characteristics of the CIFAR-100 dataset might have contributed to the suboptimal performance. These insights contribute to a deeper understanding of the interplay between model architectures and distillation processes, guiding further optimization in the field of efficient model training.

6.1. Limitations and Future Works

Our study revealed several limitations and potential areas for future research.

First, one possible reason for our model's underperformance could be the training duration. Touvron et al. demonstrated that training vision transformers can benefit significantly from extended training periods. Future work should explore running the training for a longer period to determine if the distillation benefits become more apparent over time.

Next, as we discussed in section 5.2, the CIFAR-100 dataset, with its small image size of 32x32 pixels, may not fully leverage the global contextual learning capabilities of ViTs. The added complexity of the ViT might not translate into performance gains due to the limited spatial information in small images. Future experiments could involve larger and more complex datasets, such as ImageNet, to evaluate if the distillation process yields better results when the dataset's characteristics align more closely with the strengths of ViTs.

Lastly, Touvron et al. achieved their results by thoroughly testing various configurations of CNN and ViT models. In contrast, our study focused on a single ViT and CNN model. A more comprehensive investigation involving multiple architectures and configurations might identify combinations that benefit more from the distillation process. Future work should include a systematic exploration of different ViT and CNN models to find optimal teacher-student pairs. Additionally, as discussed in 5.2, there are numerous hyperparameters that can be adjusted to potentially enhance the performance of the distilled models. These include the learning rate, batch size, weight decay, and momentum, among others. Specifically, the alpha value (the weight used in calculating the distillation loss) and the distillation type (we only utilized hard distillation) are critical parameters that warrant further experimentation. Exploring soft distillation or a combination of hard and soft distillation might yield different insights and potentially better performance. Such thorough hyperparameter tuning could uncover more effective configurations for the distillation process.

Despite the challenges encountered, our work has already shown promising results, suggesting that there is substantial potential for further advancements. We are excited to continue this line of research, confident that future work will uncover even more effective strategies for cross-architectural learning and optimization in the field of computer vision.

7. Contributions & Acknowledgment

Both authors worked together on all aspects of the paper: finding the topic, literature review, co-writing the codebase, running the models, data analysis, and writing the paper.

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[2] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers distillation through attention, 2021.

[3] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces, 2020.

[4] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[5] Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[7] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 843–852, 2017.

[8] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.

[9] Alex Krizhevsky. Learning multiple layers of features from tiny images. pages 32–33, 2009.

[10] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics Image Processing*, pages 722–729, 2008.

[11] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 554–561, 2013.

[12] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset, 2018.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[14] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks, 2019.

[15] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R. Manmatha, Mu Li, and Alexander Smola. Resnest: Split-attention networks, 2020.

[16] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers, 2020.

[17] Zhuoran Shen, Irwan Bello, Raviteja Vemulapalli, Xuhui Jia, and Ching-Hui Chen. Global self-attention networks for image recognition, 2020.

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. pages 248–255, 2009.

[19] Ross Wightman, Hugo Touvron, and Hervé Jégou. Resnet strikes back: An improved training procedure in timm, 2021.