# Knowledge Distillation for Image Classification: An Analysis of the Teacher-Student Framework

Zhong Shi
Hamilton, New Zealand

***Abstract*—This report implements and analyzes a knowledge distillation framework, transferring knowledge from a ResNet34 teacher to a ResNet18 student for image classification. The aim was to adjust the model reducing its size and maintaining accuracy. The student model, trained using a combination of cross-entropy and KL divergence losses, achieved 86.81% validation accuracy, which is 94% of the teacher's performance, but with only half the computational cost. The paper makes clear the usefulness of knowledge distillation in the establishment of effective models and addresses the gap inherent to the performance, which should be translated in terms of model capacity and logit-based transfer shortcomings, respectively, while the advanced knowledge distillation forms could be effective in the future.**

***Keywords—knowledge distillation, model compression, deep learning, image classification, ResNet, teacher-student framework***

## I. INTRODUCTION

The Convolutional Neural Networks (CNNs) and deep neural networks have turned out as being the backbone of contemporary computer vision. State-of-the-art models however tend to be big and computationally expensive imposing a limit on using these models in resource constrained devices like mobile phones and embedded systems. This will has led to the innovation of model compression and acceleration methods. Knowledge Distillation (KD) has emerged as a leading paradigm in this domain, focusing on transferring knowledge from a large, cumbersome "teacher" model to a smaller, more efficient "student" model (Hinton, Vinyals, & Dean, 2015).

The single-minded backdrop of this paper is to apply, assess and problematize the traditional knowledge distillation model. The purpose is to form a practitioner-level knowledge of its mechanisms, advantages, and the limitations therein. We shall train a ResNet34 as the teacher model and distill its knowledge to a non-pretrained ResNet18 student model on standard at the image classification task. In this report, the whole process of work is going to be described data preprocessing, model training, and evaluation. It will provide analysis of the performance disconnect between the teacher and the distilled student and using recent academic literature, it will deconstruct the reasons behind the same.

## II. LITERATURE REVIEW

The training of knowledge distillation was formalized by Hinton, Vinyals, & Dean (2015), who offered to compress the knowledge of a large model to a small one. The core idea is that the softened probability distribution produced by a teacher model contains richer information about the similarity between classes—termed "dark knowledge"—than the one-hot hard labels used in standard training. Training the student model then proceeds by training it to reproduce this soft distribution by trying to minimize a Kullback-Leibler (KL) divergence.

The teacher model is the most important thing, basing the maximum delivery of knowledge. There is the quality of such knowledge that is determined by the process of training the teacher. Szegedy et al. (2016) identified that training on hard labels can lead to over-fitting and make the model "too confident about its predictions." They proposed a method of smoothing labels during the model training, Label Smoothing Regularization (LSR), which may help promote the better model confidence and, consequently, better generalization and possibly better soft labels to be distilled.

Moreover, it is important to select a model of a teacher. In their research on Data-efficient Image Transformers (DeiT), Touvron et al. (2021) found something remarkable. They discovered that, in the case of a Vision Transformer (ViT) student, a Convolutional Neural Network (CNN) teacher may usually produce good distillation outcomes compared to another ViT teacher. This is attributed to the CNN's ability to transfer its inherent inductive biases (e.g., locality, translation invariance) to the architecturally different student. This highlights that the teacher's architecture and the knowledge it embodies are key factors in a successful distillation process.

The previous studies strictly analyzed how far the traditional KL divergence loss goes against the limitations. In the case of the KD loss, Zhao et al. (2022) reformulated it and unveiled it as a coupled formulation. Their work shows that the knowledge from non-target classes (the core of "dark knowledge") is suppressed when the teacher is highly confident in its prediction for the target class. They suggested a Decoupled Knowledge Distillation (DKD) that decoupled the learning of target-class knowledge and non-target-class knowledge, and enable more and flexible knowledge transfer.

On this critique, to add, another base failure of the conventional KD setup was noted by Sun et al. (2024). The shared temperature parameter implicitly forces the student's logits to match the teacher's logits in both numerical range and variance. This is an expensive limitation to a smaller student model. To separate the learning of logit relations with the necessity to align their magnitude, they suggested a simple preprocessing scheme of Logit Standardization, a simple Z-score scheme.

Lastly, the conversion of knowledge does not involve the last output level only. This lack of the sufficiency of logits use alone preconditioned the body of related work concerned with feature-based distillation. This was specifically considered by Yang et al. (2024) in ViTKD paper in the case of Vision Transformers. They discovered that direct feature copying commonly done by CNNs can harm ViTs performance. Their analysis revealed that different layers in a ViT require distinct distillation strategies, proposing "mimicking for shallow layers and generation for deep layers," which underscores the need for architecture-aware distillation methods.

## III. Methodology

Our methodology was scheduled in the form of a three-stage pipeline that iteratively would carry out the process of knowledge distillation and caliber its performance. The initial step entailed the thorough data preprocessing procedure. They loaded images contained in their respective class folders and subject these to a uniform transformation process to make them suitable and applicable in training the network. The first step is to resize all the images to constrain their aspect ratio, but scaling the smallest side of each of them to 224 pixels, as well as a center cropping to reach a final size of 224x224. Subsequently, having normalized pixel values to a floating-point range of, mean and standard deviation values at a per-channel level were computed upon the entire training set. These numbers were then used to normalize all the datasets (training, validation, and test) which is an important process of stabilizing the training process of deep networks. Lastly, data format was transformed to the form needed in PyTorch NCHW tensor.

The second phase was devoted to the preparation of teacher model. We picked a ResNet34, which would give us a solid base with powerful pre-trained weights based on ImageNet. The last layer of classifications was substituted with a new one and fitted with the 10 classes of our unique data. This model was thereafter tuned on own training data preprocessed. The training was prompted by the standard cross-entropy loss and Adam optimizer. In the process, we added a tracker of performance on the validation set and saved the weights of the model that reached the highest validation accuracy. This made our teacher the epitome of knowledge that could be derived out of the information utilising this architecture.

The last and the most crucial phase was the use of the knowledge distillation process. In the case of student model we selected ResNet18 architecture, but deliberately it was not pre-trained, so it had to learn completely during distillation operation. The training of the student was done via a composite loss. This defeat is thoughtfully balanced set of two different messages of learning. The first is the conventional cross-entropy loss, calculated between the student's raw predictions and the ground-truth hard labels, which anchors the student's learning to the factual data. The KL divergence loss is the second one, but it is central to distillation. This loss computes the variation between the smoothed probability distributions of the student and teacher models.

## IV. Results

After undergoing its fine-tuning procedure throughout its 20 epoch periods, the teacher model (ResNet34) created a high-performance norm by attaining the test recognition rate of 92.35% during its peak period. With this good performance we were assured of getting a good source of knowledge when guiding the student. Then the ResNet18 student model was trained using distillation framework in scratch 60 epochs. Its accuracy during peak validation is 86.81 percent that occurred at epoch 33. After this stage, it started performing more or less at the same level on the validation set, and then with a gradual decrease, which suggests that additional training would not have brought substantially beneficial results and would have resulted in overfitting. This fact led to applying an early stopping method, which ended the training at epoch 43.

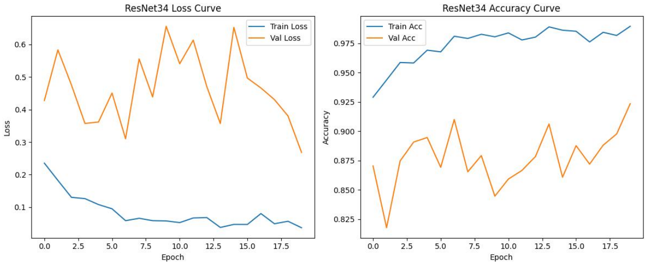

Fig. 1. Training and Validation Curves for the ResNet34 Teacher Model.

Examination of the training curves is an additional source of information about the dynamics of learning. The teacher model's loss and accuracy curves demonstrated a smooth and stable convergence, as expected from a fine-tuning process. The student model's validation accuracy curve, presented below, reveals a steep and rapid ascent during the initial epochs, a testament to the powerful guidance provided by the teacher. The curve then gracefully plateaus in the 86-87% range, a level considerably higher than what a non-distilled ResNet18 would typically achieve, yet visibly below the teacher's 92% ceiling. In this visual representation, the difference between the number of knowledge transfer success and the performance left is obvious.

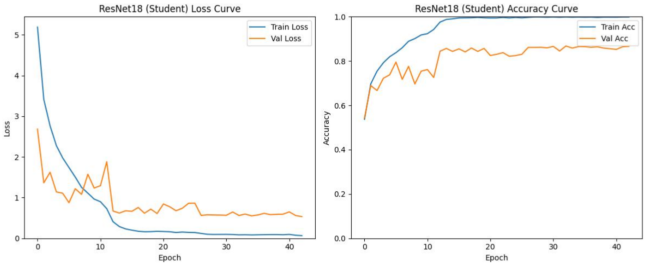

Fig. 2. Training and Validation Curves for the ResNet18 Student Model under Knowledge Distillation.

Comparing the two in absolute terms, there is a 5.54 percent gap between the teacher and the most proficient student model. While the student did not fully match the teacher's performance, the primary goal of distillation was achieved in the trade-off between accuracy and efficiency. Approximately 11.7 million-parameter student has near to half the computational cost (1.8 GFLOPs) of 21.8 million-parameter teacher that requires 3.6 GFLOPs. In essence, the student model successfully delivered 94% of the teacher's accuracy while demanding only 50-54% of its computational and parameter budget, highlighting a highly effective and practical application of model compression.

## V. Discussion

The experimental outcome provided a convincing argument on the classic knowledge distillation technique as a practical tool of building powerful models that are efficient. The ResNet18 student reached a final accuracy of 86.81%, which is a great success of such a small model trained end to end. This outcome confirms that the "dark knowledge" transferred from the ResNet34 teacher provides a rich and effective learning signal. The analysis should not however end at this success, the rather tell-tail 5.54 percent performance discrepancy between the teacher and the student is worth further investigation, which when considered through the prism of modern research, should be the limitation of this canonical distillation practice.

The simplest explanation of such performance difference is the intrinsic difference in capacity between the ResNet34 and ResNet18 architectures. The student model is just fewer parameters and shallower architecture, which reduces its capacity to apply those complex functions it has picked up from the more powerful teacher as perfectly.

Besides, the nature and quality of transference of knowledge is vital. The teacher's "knowledge" is itself a product of its training on hard, one-hot labels. The training process also brings about too confident predictions and does not explicitly encode a rich relationship between classes. As Szegedy et al. (2016) noted, this encourages the model to become "too confident," potentially limiting the quality of the "dark knowledge" it can provide.

It is not the process of distillation that is actually a problem, but some of its components. By relying solely on matching the final logits, we discard the vast amount of information contained in the teacher's intermediate feature maps. That is like a student being presented with the final answer to a mathematical problem and not the procedure he/she went through to reach that answer. The drawbacks of the traditional KL divergence loss have also been demonstrated in recent work, such as DKD, and it was shown that such an objective critically intertwines the acquisition of target and non-target class information, and that it saturates the signal of the classical high-confidence samples (Zhao et al., 2022). Additionally, the work on Logit Standardization empirically proves that forcing the student to match the teacher's logit magnitudes is an unnecessary and often harmful constraint (Sun et al., 2024).

## VI. Conclusion

This research was able to establish a basic knowledge distillation scheme whereby we see that it has the ability to achieve a student model that provides computational efficiency and a lot of performance. By transferring knowledge from a ResNet34 teacher, our ResNet18 student achieved a remarkable balance, securing 94% of the teacher's accuracy while requiring only half the computational resources. This finding confirms the main claim of knowledge distillation of being a highly effective tool of model compression.

But as we have analyzed and as it is in tune with modern studies, the process is not over yet. The constant gap between the results seen on the performance of the teacher and the student is not the failure of the process but the solid sign of its limits. We identified three primary frontiers for advancement: the quality of the teacher's initial knowledge, the efficiency of the knowledge transfer mechanism, and the student's own capacity to learn.

## References

[1] Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the Knowledge in a Neural Network. arXiv preprint arXiv:1503.02531.

[2] Sun, S., Ren, W., Li, J., Wang, R., & Cao, X. (2024). Logit Standardization in Knowledge Distillation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[3] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the Inception Architecture for Computer Vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[4] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. In International Conference on Machine Learning (ICML).

[5] Yang, Z., Li, Z., Zeng, A., Li, Z., Yuan, C., & Li, Y. (2024). ViTKD: Feature-based Knowledge Distillation for Vision Transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops.

[6] Zhao, B., Cui, Q., Song, R., Qiu, Y., & Liang, J. (2022). Decoupled Knowledge Distillation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).