Enhanced Knowledge Distillation using Intermediate Neural Networks

Tasin Towsif Rahman
Department of ECE
North South University
Dhaka, Bangladesh
tasin.towsif@northsouth.edu
1911774042

Fahim Faysal Apurba
Department of ECE
North South University
Dhaka, Bangladesh
fahim.apurba@northsouth.edu
1921442642

Pious Alam Patwary
Department of ECE
North South University
Dhaka, Bangladesh
pious.patwary@northsouth.edu
1931459042

MD Sabbir Hossain
Department of ECE
North South University
Dhaka, Bangladesh
sabbir.hossain02@northsouth.edu
1610443042

***Abstract*—Deep neural networks are effective models that perform well on a variety of tasks, but they are too big to be used on devices like smartphones or smaller embedded systems. By compressing a potent but heavy pre-trained neural network (teacher) into a lightweight neural network (student), knowledge distillation seeks consistent performance, where the goal is to have the student model mimic the behavior and predictions of the teacher model. Over the years, there have been several approaches to improve the distillation of knowledge from the large teacher model to the smaller student model, such as self-distillation, ensemble distillation, attention transfer, teacher-assistant distillation, etc. In this paper, we show that if the gap between the teacher model and the student model is too large, the performance of knowledge distillation falls. Thus, we engage an intermediate neural network (teacher-assistant/TA) to bridge the gap between the teacher and the student model and reuse the teacher model's fully-connected layer in the student model. Moreover, we study the effects of using various teacher and teacher-assistant models for knowledge distillation on a fixed student model, both with and without reusing the teacher's fully-connected layers through experiments on the CIFAR-10 and CIFAR-100 datasets using various ResNet architectures.**

***Index Terms*—knowledge distillation, teacher-student distillation, distillation, model compression**

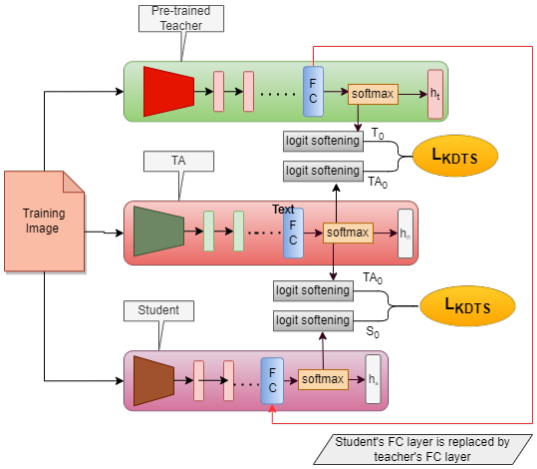

Fig. 1. An illustration of the the full overview of our proposed work. Here, T_0, TA_0, S_0 are the soft outputs of the teacher, TA, and the student model respectively. FC refers to the fully connected layers of the respective models. In our work, we reuse the fully connected layer of the teacher model in the student model to help it perform just as well as the teacher during inference time.

I. Introduction

In the realm of deep learning, the idea of knowledge distillation—first proposed by Hinton et al. [1]—has become a potent method. It entails the transmission of information from a larger, more complicated model (the teacher) to a smaller, simpler model (the student) [2]. This strategy's main goal is to capitalize on big models' predictive potential while reducing their computational and memory requirements, which makes them more deployable on devices with limited resources [3].

Instead of simply reiterating the final class predictions, the knowledge distillation process involves training the student model to mirror the output distribution of the teacher model [4]. By allowing the student model to absorb the teacher model's extensive representational information, generalization performance is enhanced [5].

Despite its promise, knowledge distillation is not without challenges. One of the key issues is the selection of an appropriate teacher model. The teacher model needs to be sufficiently complex to capture the underlying data distribution, yet simple enough to facilitate efficient knowledge transfer [6]. Another challenge is the design of an effective distillation loss function, which should encourage the student model to learn both the hard targets (i.e., the ground-truth labels) and the soft targets (i.e., the teacher's output distribution) [7].

However, the gap in performance between the distilled student and the original teacher model is still significant. To tackle this drawback, numerous approaches have been introduced over the past few years [8, 9]. The majority of them take advantage of leveraging the additional oversight provided by the pre-trained instructor model, especially the intermediate layers [10, 11, 12, 13, 14, 15, 16]. Despite the fact that these works do really help students perform better over time, neither effective representations nor optimal hyperparameters that guarantee their success are simple to implement in actual practice. A consistent and precise understanding of the ultimate improvement in student performance is further impeded by the variability of imparted knowledge.

Inspired by this observation, Mirzadeh et al. [17] proposed

and worked on a new distillation framework called Teacher Assistant Knowledge Distillation, which introduces intermediate neural networks (teacher-assistant/TA) between the teacher and student to try and bridge the gap between the two. Chen et al. [18] worked on improving the capability of the student model by reusing the teacher classifier for better generalization of the outputs.

In this paper, we leverage the concepts introduced to run a series of experiments in an attempt to narrow the rift regarding performance and accuracy after knowledge distillation compared to the teacher. We implement the TAKD framework, where there is a pre-trained teacher model, a teacher-assistant/TA model, and a student model. Knowledge distillation occurs from the teacher to the TA to the student model. The student model reuses the fully connected layer/classifier layer of the teacher model to help better generalize the outputs, in addition to utilizing the dark knowledge obtained from the TA model by distillation. An illustration of our proposed idea is demonstrated in **Fig. 1**.

II. Related Work

Hinton et al. [1] first presented the idea of moving knowledge from a larger model (teacher) to a smaller model (student), which is how the concept of knowledge distillation came to be. Since then, a vast amount of deep learning research has been motivated by this important work. Bucilu et al.'s [2] introduction of the notion of model compression—which entails shrinking the instructor model while maintaining its predictive power—further investigated this issue. This concept was developed by Ba and Caruana [3], who showed that shallow models can perform as well as deep models when they are taught to replicate their output.

Several strategies have been proposed to enhance the effectiveness of knowledge distillation. For instance, Romero et al. [12] introduced the concept of "FitNets", which involves training the student model to mimic the intermediate representations of the teacher model. This approach has been shown to be particularly effective for training compact models that achieve high performance on complex tasks [19].

Lopez-Paz et al. [20] proposed the idea of "Dark Knowledge", which involves training the student model to learn from the teacher's mistakes. This approach has been shown to be effective for improving the robustness of the student model against adversarial attacks [21].

Zagoruyko and Komodakis [6] introduced the concept of "attention transfer", which involves training the student model to mimic the attention maps of the teacher model. This approach has been shown to be effective for improving the interpretability of the student model [22].

Yim et al. [7] proposed the idea of "knowledge distillation with fast optimization", which involves training the student model to learn both the hard targets (i.e., the ground-truth labels) and the soft targets (i.e., the teacher's output distribution) simultaneously. This approach has been shown to be effective for accelerating the training process of the student model [23].

Several other variants of knowledge distillation have also been proposed. For instance, Park et al. [24] introduced the concept of "relational knowledge distillation", which involves training the student model to mimic the relational knowledge of the teacher model. Meanwhile, Kim et al. [25] proposed the idea of "self-distillation", which involves training the model to distill its own knowledge.

Despite the significant progress in the field of knowledge distillation, there are still several open challenges that need to be addressed. One of the key issues is the selection of an appropriate teacher model. The teacher model needs to be sufficiently complex to capture the underlying data distribution, yet simple enough to facilitate efficient knowledge transfer [26]. Another challenge is the design of an effective distillation loss function, which should encourage the student model to learn both the hard targets and the soft targets [27].

A new distillation framework known as Teacher Assistant Knowledge Distillation was designed and tested by Mirzadeh et al. [17]. This framework places intermediate neural networks (teacher-assistant/TA) between the teacher and student in an effort to close the communication gap between the two. By leveraging the teacher classifier to improve the generalization of the outputs, Chen et al.'s [18] research aimed to enhance the student model's performance. Our work leverages the best of both aimed towards improving the distillation of knowledge from teacher to student model while keeping the generalization abilities of the teacher model as intact as possible.

III. Proposed Methodology

There are many knowledge distillation techniques that have been proposed recently for reducing the knowledge gap between teacher and student. In most cases, a direct teacher-to-student knowledge distillation approach is used. But from a big teacher model to a very small student model, knowledge distillation is very costly, and a big gap in performance between teacher and student has been found. And our goal is to reduce that gap. This is why, we distilled knowledge from Teacher to TA first, then, TA to the student, where the teacher's fully-connected layer has been reused in the student model as well.

A. Dataset:

a. First, we obtain the CIFAR-10 and CIFAR-100 datasets, which consist of 32x32 color images across 10 and 100 classes, respectively.
b. Then we split each dataset into training, validation, and test sets, where batch size was 128 for each case.

B. Teacher Model Training

a. We trained Resnet32 as the teacher model using the CIFAR-10 and CIFAR-100 training sets.
b. Evaluate the teacher model's performance on the CIFAR-10 and CIFAR-100 training and test sets to establish a baseline.

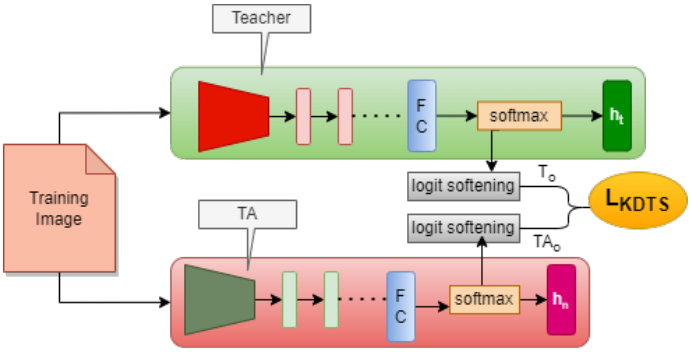

Fig. 2. Teacher to TA Knowledge distillation.

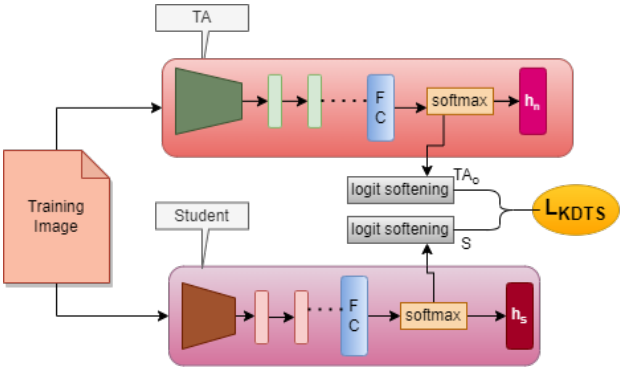

Fig. 3. TA to Student Knowledge distillation using base student model.

C. Teacher-to-TA Knowledge Distillation

a. We chose Resnet20 as the TA model, which is a smaller and more lightweight model compared to the teacher model.
b. Then distill the TA model from the teacher model on CIFAR-10 and CIFAR-100 training sets.
c. Minimize the distillation loss between the TA's predictions and the soft targets from the teacher model.
d. Here, we use a temperature value as T = 3.16 and formulate a loss function like,

$$L_{\text{KDTS}} = \text{KLDiv}\left(\text{softmax}\left(\frac{\text{TAo}}{T}\right), \text{softmax}\left(\frac{\text{To}}{T}\right)\right) \cdot T^2$$

,
where, TAo = TA's soft outputs and To =Teacher's soft outputs.
We have used this KDTS loss insted of this simple KD loss

$$L = \text{KLDiv}\left(\text{softmax}\left(\frac{\text{TAo}}{T}\right), \text{softmax}\left(\frac{\text{To}}{T}\right)\right)$$

e. Then, the TA model's performance has been evaluated on the CIFAR-10 and CIFAR-100 train and test set and compared it against the teacher model.

D. TA-to-Student Knowledge Distillation with Reuse of Fully-Connected Layer:

a. We selected Resnet8 as student model, which was smaller or shallower than both the teacher model and the TA model.
b. The fully connected layer of the Teacher model has been reused in the student model.
c. Distill the student from the TA model on the CIFAR-10 and CIFAR-100 training sets.
d. Minimize the distillation loss between the student's predictions and the soft targets from the TA model.
e. Again, using the same temperature value T=3.16, and the KDTS loss function,

$$L_{\text{KDTS}} = \text{KLDiv}\left(\text{softmax}\left(\frac{\text{So}}{T}\right), \text{softmax}\left(\frac{\text{TAo}}{T}\right)\right) \cdot T^2$$

,
where, So = Student's soft outputs and TAo =TA's soft outputs.
f. Evaluate the student model's performance on the CIFAR-10 and CIFAR-100 test sets and compare it against the teacher model and TA model.

IV. Experimental Setup

a. We trained Resnet32 as the teacher model using the CIFAR-10 and CIFAR-100 training sets.
b. For training, the optimization algorithm used was stochastic gradient descent (SGD) or Adam.
c. Evaluate the teacher model's performance on the CIFAR-10 and CIFAR-100 training and test sets to establish a baseline.

A. Dataset Selection

- We have chosen CIFAR-10 and CIFAR-100 datasets for the knowledge distillation experiments.
- CIFAR-10 consists of 50,000 training images and 10,000 test images, while CIFAR-100 contains 50,000 training images and 10,000 test images.
- We have ensured that the datasets are preprocessed and transformed appropriately, such as normalizing pixel values, where the training set consists of 80 percent of data and the test set consists of 20 percent of data.

B. Model Architectures

- We have selected Resnet32, Resnet20 and Resnet8 model architectures for accordingly the teacher, TA, and student models.

C. Hyperparameter Tuning

- We have determined the hyperparameters values relevant to the knowledge distillation process, such as temperature($T = 3.16$) for the softmax operation and a learning rate of 0.01 for optimization.
- Hyperparameter tuning has been performed using validation sets to find optimal values that enhance knowledge transfer.

D. Training Procedure

- The teacher model has been trained using the CIFAR-10 and CIFAR-100 training sets until convergence or for a specific number of epochs.
- Soft targets have been generated using the teacher model over the training set by dividing the softmax outputs by the temperature (T).
- TA model has been distilled from the teacher model on the training set while minimizing the Kullback-Leibler divergence (KLDiv) loss between the TA's and teacher model's soft targets, which provides the TA model with the 'dark knowledge' to help it perform better during inference time. **Fig. 2** shows a representation of this process
- The same steps have been repeated for knowledge distillation between the TA and the student model.

E. Evaluation Metrics

- The performance of the teacher, TA, and student models have been evaluated on the CIFAR-10 and CIFAR-100 test sets, based on the percentage accuracy of their predictions.
- The performance of the teacher model, TA model, and student model has been observed individually as well to establish the benefits of knowledge distillation.
- Improvements have been analyzed in accuracy, model size, and computational efficiency achieved through the distillation process.

F. Reproducibility

- All experimental settings, including dataset details, model architectures, hyperparameters, and training procedures have been documented to ensure reproducibility.

V. Results and Analysis

First, the teacher model Resnet32 is trained on the CIFAR-10 dataset, and after 65 epochs it reached an accuracy of 81.11 percent. Then, we have distilled knowledge to the TA model, Resnet20, while employing the KDTS loss function.
In the same way, on the CIFAR-100 dataset, the teacher model is trained and after 65 epochs it reached an accuracy of 51.41 percent. Then, we have distilled knowledge to the TA model
We experiment the **Teacher** to **TA** to **Student** knowledge distillation in 4 different ways on the basis of Knowledge distillation loss function and reusing the teacher's FC layer in student.
Here, is the 4 different combination below:
1. For training student, using simple KD loss + base student model
2. For training student, using simple KD loss + reuse teacher's FC in sutdent model
3. For training student, using KDTS loss + base student model (**Fig. 3** shows a representation for this.)
4. For training student, using KDTS loss + reuse teacher's FC in sutdent model.

After experiment we get:

TABLE I
Performance Comparison Table For Cifar-10 dataset

	Performance (Accuracy)			
	Teacher	**TA**	**Base student**	**Student reuse FC**
L=KDTS	**81.11**	**79.33**	**75.82**	**76.94**
L=KD	**81.11**	**71.31**	**69.76**	**72.54**

From the table-I, we can say that for **TA** to **Student** knowledge distillation, (4. using KDTS loss + reuse teacher's FC in student model) combination has produced the best performance results on the CIFAR-10 dataset.

TABLE II
Performance Comparison Table For Cifar-100 dataset

	Performance (Accuracy)			
	Teacher	**TA**	**Base student**	**Student reuse FC**
L=KDTS	**51.41**	**48.65**	**41.62**	**42.58**
L=KD	**51.41**	**43.12**	**33.39**	**34.35**

From both tables, we can observe and say that for **TA** to **Student** knowledge distillation, (4.using Our Special KDTS loss + reuse teacher's FC in sutdent) combination has been produced best performance on CIFAR-10 dataset.

So, in conclusion, it is to say that in distillation of knowledge Teacher to TA and TA to Student, when, teacher's FC layer is reused in student and the knowledge distillation loss is KDTS, which is- L_{KDTS} = KLDiv $\left(\text{softmax}\left(\frac{\text{So}}{T}\right), \text{softmax}\left(\frac{\text{TAo}}{T}\right)\right) \cdot T^2$, is applied, knowledge gap between teacher and student reaches minimum.

Acknowledgment

We would like to convey our sincere gratitude to Dr. Nabeel Mohammed, our course instructor, for his guidance and support during the project.

References

[1] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," arXiv preprint arXiv:1503.02531, 2015.
[2] C. Buciluǎ, R. Caruana, and A. Niculescu-Mizil, "Model compression," in Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, 2006, pp. 535-541.
[3] J. Ba and R. Caruana, "Do deep nets really need to be deep?," in Advances in neural information processing systems, 2014, pp. 2654-2662.
[4] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," nature, vol. 521, no. 7553, pp. 436-444, 2015.
[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, 2012, pp. 1097-1105.

[6] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," arXiv preprint arXiv:1612.03928, 2016.

[7] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4133-4141.

[8] Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. Knowledge distillation: A survey. Int. J. Comput. Vis., 129(6):1789–1819, 2021.

[9] Lin Wang and Kuk-Jin Yoon. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. IEEE Trans. Pattern Anal. Mach. Intell., 2021.

[10] Sungsoo Ahn, Shell Xu Hu, Andreas C. Damianou, Neil D. Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9163–9171, 2019.

[11] Defang Chen, Jian-Ping Mei, Yuan Zhang, Can Wang, Zhe Wang, Yan Feng, and Chun Chen. Cross-layer distillation with semantic calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7028–7036, 2021

[12] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *International Conference on Learning Representations*, 2015.

[13] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations*, 2020.

[14] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *International Conference on Computer Vision*, pages 1365–1374, 2019.

[15] Jing Yang, Brais Martínez, Adrian Bulat, and Georgios Tzimiropoulos. Knowledge distillation via softmax regression representation learning. In *International Conference on Learning Representations*, 2021.

[16] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. In *International Conference on Learning Representations*, 2017.

[17] S. I. Mirzadeh, M. Farajtabar, A. Li, N. Levine, A. Matsukawa, and H. Ghasemzadeh, "Improved Knowledge Distillation via Teacher Assistant," Proceedings of the *AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, pp. 5191–5198, 2020.

[18] Defang Chen, Jian-Ping Mei, Hailin Zhang, Can Wang, Zhe Wang, Yan Feng, and Chun Chen. 'Knowledge distillation with the reused teacher classifier', IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022 [Preprint].

[19] Y. Chen, Y. Wang, Z. Kira, and Y. Jia, "DarkRank: Accelerating Deep Metric Learning via Cross Sample Similarities Transfer," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3017-3025.

[20] D. Lopez-Paz, L. Bottou, B. Schölkopf, and V. Vapnik, "Unifying distillation and privileged information," in *Proceedings of the International Conference on Learning Representations*, 2016.

[21] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks," in 2016 IEEE Symposium on Security and Privacy (SP), 2016, pp. 582-597.

[22] W. Xu, Y. Sun, X. Deng, and H. Zhang, "Attention Transfer from Web Images for Video Recognition," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 1790-1798.

[23] J. H. Park, S. Woo, J. Y. Lee, and I. S. Kweon, "BAM: Bottleneck Attention Module," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

[24] Y. Park, D. Kim, G. Kim, and S. Lee, "Relational Knowledge Distillation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3967-3976.

[25] J. Kim, S. Park, and N. Kwak, "Paraphrasing Complex Network: Network Compression via Factor Transfer," in Advances in Neural Information Processing Systems, 2018, pp. 7315-7325.

[26] Q. Xie, M. Luong, E. Hovy, and Q. V. Le, "Self-training with Noisy Student improves ImageNet classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10687-10698.

[27] Y. Tian, D. Krishnan, and P. Isola, "Contrastive Representation Distillation," in *Proceedings of the International Conference on Learning Representations*, 2020.