Improving Knowledge Distillation Using Data Augmentation

Jun Seung Woo, Shin Dong Ho,
Student and Professor, MY PAUL SCHOOL
12-11, Dowontongmi-gil, Cheongcheon-myeon, Goesan-gun,
Chungcheongbuk-do, Republic of Korea
eavatar@hanmail.net

Abstract

Knowledge Distillation (KD) is a type of transfer learning where a large, pre-trained model (Teacher Model) transfers knowledge to a smaller, simpler model (Student Model) to make the latter more efficiently trainable. Knowledge distillation (KD) can effectively transfer the knowledge of a teacher model to a student model in image classification tasks. However, the results of KD can vary depending on the specific method and approach used, thus a standardized improvement method is necessary. Data Augmentation techniques are powerful performance improvement methods for image classification. In this paper, we propose a standardized approach to improving knowledge distillation using various Data Augmentation techniques based on the α value. The results showed that Data Augmentation techniques provided a significant performance improvement in the Knowledge Distillation model. We demonstrated that as the strength of data augmentation increases, the performance of knowledge distillation (KD) also increases proportionally. And It seems that the α value does not have a significant impact on the performance improvement of knowledge distillation (KD).

Keywords
deep learning, computer vision, Image Classification, Knowledge Distillation and Data Augmentation.

1. Introduction
In deep learning, transfer learning refers to the process of transferring the knowledge learned by one model to another model that has not been trained. This is typically achieved through a concept called knowledge distillation (KD), which is designed to transfer the knowledge learned by a large, pre-trained model (teacher model) to a smaller model (student model) to improve the efficiency of the student model learning.

Currently, deep learning models are becoming deeper and wider, with more parameters and computational requirements, and as a result, the performance of these models continues to improve. Knowledge distillation aims to enable smaller networks to achieve similar performance to larger networks, as shown in Figure 1, by transferring the knowledge learned by a pre-trained teacher model to the student model. Many papers have demonstrated that using knowledge distillation during the learning process can improve performance and stability (Ridnik et al. 2022)(Gou et al. 2021)(Cho et al. 2019). In addition, there have been several attempts to improve the performance of Knowledge Distillation (KD) by applying data augmentation separately to KD, as demonstrated by (Fu et al. 2020) and (Wang et al. 2019), and by enhancing the Teacher model used in KD, as shown in various attempts such as (Mirzadeh et al. 2020) and (Beyer et al. 2022). Other methods have improved Knowledge Distillation (KD) by varying the ways in which knowledge is distilled such as response-based distillation (Hinton et al. 2015)(Feng et al. 2021), feature-based distillation (Park et al. 2019)(Heo et al. 2019) and relation-based distillation (Romero et al. 2014).

Especially in (Yim et al. 2017). The authors demonstrate the effectiveness of KD in improving the speed and accuracy of model optimization, reducing the size of deep neural networks, and facilitating transfer learning between different tasks and domains. The authors evaluate the effectiveness of KD on various benchmark datasets and tasks, including image classification, object detection, and natural language processing. They demonstrate that KD can significantly improve the performance of deep learning models while reducing their computational complexity and memory requirements.

However, these various methods are merely ways to expect effective optimization performance for KD in different environments, and cannot be considered standardized improvement methods in an integrated environment.

This paper aims to apply data augmentation, a powerful performance enhancement technique in computer vision, to knowledge distillation and to further improve the performance of knowledge distillation in various integrated environments.

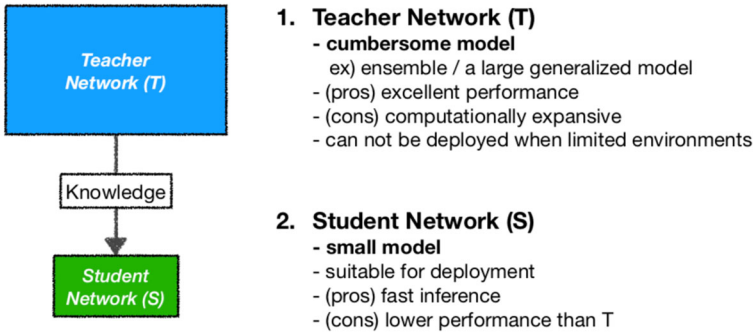

Figure 1. Knowledge Distillation Structure

2. Body

2.1 Knowledge Distillation

Knowledge Distillation (KD) is a type of transfer learning in deep learning where a pre-trained model transfers its knowledge to another model that has not yet been fully trained, allowing the latter to learn more efficiently than it would otherwise. The idea of KD was first proposed in 2014 by Hinton. KD is typically used to train smaller models with fewer parameters and are less deep than larger models, referred to as the Teacher Model and the Student Model, respectively. The logit values produced by the Teacher Model are used to train the Student Model, aiming to improve the latter's performance.

Hinton's KD approach utilizes two main methods: Soft Labels and Distillation Loss. Soft Labels and Distillation Loss. Soft Labels involve using a hyperparameter called Temperature (T) to soften the original hard labels, making them easier to transfer to the Student model. In a traditional deep learning model, the output of the Softmax function (i.e., $\exp(z_i)$) is either close to 1 if the predicted label is correct, or close to 0 if it is incorrect. This can make the output very extreme and not very informative for the Student model to learn from. However, Soft Labels divide $\exp(z_i)$by the Temperature parameter, as shown in Equation 1 in the original paper (Hinton et al. 2015), allowing even small values of $\exp(z_i)$ to have a significant impact on the predicted label. By doing so, Soft Labels enable the Teacher model to pass on more nuanced information to the Student model, in the form of $\exp(z_i/T)$ values, depending on the value of T. When T is low, the values are similar to traditional probability values (i.e., close to 0 or 1), whereas when T is high, the values are more softly distributed, making it easier for the Student model to learn from the teacher's output.

$$q_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_i/T)}$$

Equation 1. Hinton's Soft Label

Distillation Loss, as proposed in Hinton et al. 2015, uses KL Divergence to compute a loss based on the Soft Labels produced by the Teacher Model. The Soft Labels are used as a new set of labels for the Student Model to learn from, in addition to the ground truth labels from the training data. As shown in Figure 2, the Student Model receives both the Soft Labels from the Teacher Model and the ground truth labels and computes a loss function based on these labels. The KD loss L_{KD} is then computed using KL Divergence between the Soft Labels produced by the Teacher Model (q_i) and the Soft Labels produced by the Student Model (p_i), as shown in Equation 2. The Student Loss L_{CE} is the standard Cross-Entropy (CE) loss, which compares the Student Model's Hard predictions with the ground truth labels. The final Loss is computed using a parameter α that controls the strength of L_{KD} and L_{CE}.

$$L_{KD} = \sum_{k} D_{KL}(q_k \,||\, p_k)$$
$$Loss = \alpha L_{KD} + (1-\alpha) L_{CE}$$

Equation 2. Hinton's Distillation Loss

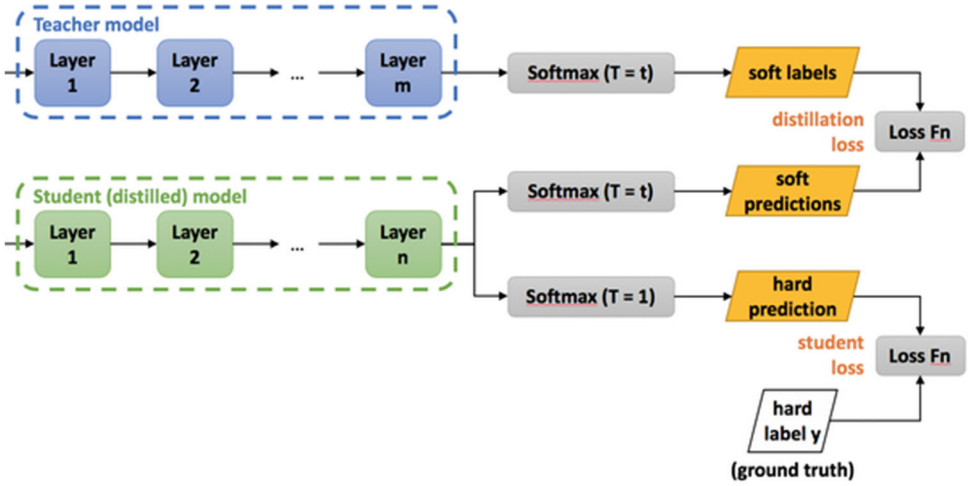

Figure 2. Hinton's Knowledge Distillation Structure

2.2 Data Augmentation

Data augmentation is a technique that increases the effective size of a training dataset by augmenting the existing data through various methods. Data augmentation has been a powerful way to improve the performance of models in computer vision, and research has been conducted to identify and efficiently use various data augmentation methods. In this study, the data augmentation techniques used were RandomCrop, RandomHorizontalFlip, and RandAugment (Cubuk et al. 2020). RandomCrop randomly cuts the image to perform augmentation based on the specified size, while RandomHorizontalFlip horizontally flips the input image, randomly performing left-right inversion for augmentation. Lastly, RandAugment applies multiple data augmentations simultaneously, using two parameter values, N and M, where N is the number of augmentations to apply, and M represents the strength of the augmentations. With the N and M values found, multiple augmentations are randomly applied using RandAugment. These data augmentation methods have shown good results in most tasks and were thus chosen for this study's experiments.

2.3 Experiment Results

In this study, we experimented with applying data augmentation, which has shown good performance in the computer vision field, to image classification using Knowledge Distillation (KD). We also experimented with various data augmentation techniques based on the α values in the distillation loss, to investigate whether KD can benefit from data augmentation. We used the well-known CIFAR100 dataset for the training data, and ResNet32 and ResNet8 for the teacher and student models, respectively. We experimented with five different α values for the temperature (T) hyperparameter value, which was set to 4. We used augmentation techniques such as RandomCrop, RandomHorizontalFlip and RandAugment, which have been shown to be effective in image classification, and set the N and M values of RandAugment to Cubuk et al. 2020, Park et al. 2019, which showed the best results for WideResNet in a previous study (Zagoruyko et al. 2016).

In this study, we categorized data augmentation into three groups: non, RandomHorizontalFlip + RandomCrop, and RandomHorizontalFlip + RandomCrop + RandAugment, and observed the results as the augmentations became stronger. We also used a linearly increasing method to set α values at 0, 0.25, 0.5, 0.75, and 1.0, and intuitively observed the results when data augmentation and α values linearly increased.

The experimental results confirmed that applying data augmentation to models learned through KD resulted in an average accuracy increase of more than 14% compared to models without augmentation as shown in Figure 3. We also found that the α parameter, which adjusts the intensity of the distillation loss, did not have a significant effect on accuracy improvement when using data augmentation. However, an interesting phenomenon was observed where performance improvement was relatively lower when α was set to 1.0, when learning using L_{KD} without using L_{CE} (Table 1).

Through our experiments, we found that data augmentation can significantly improve the performance of knowledge distillation (KD) for image classification problems. As we applied more powerful data augmentation methods, we observed a consistent improvement in KD performance. These results suggest that further research into more advanced data augmentation techniques may lead to even better performance in KD. However, the α parameter does not significantly affect accuracy improvement when using data augmentation.

Table 1. Experiment Results

alpha	**0**	**0.25**	**0.5**	**0.75**	**1.0**
Non	59.32	58.92	59.5	59.44	59.87
Flip+crop	72.39	73.44	73.26	72.83	71.56
Flip+crop+RA	73.59	73.93	74.32	73.07	71.53
Gains	**+14.27**	**+15.01**	**+14.82**	**+13.63**	**+11.66**

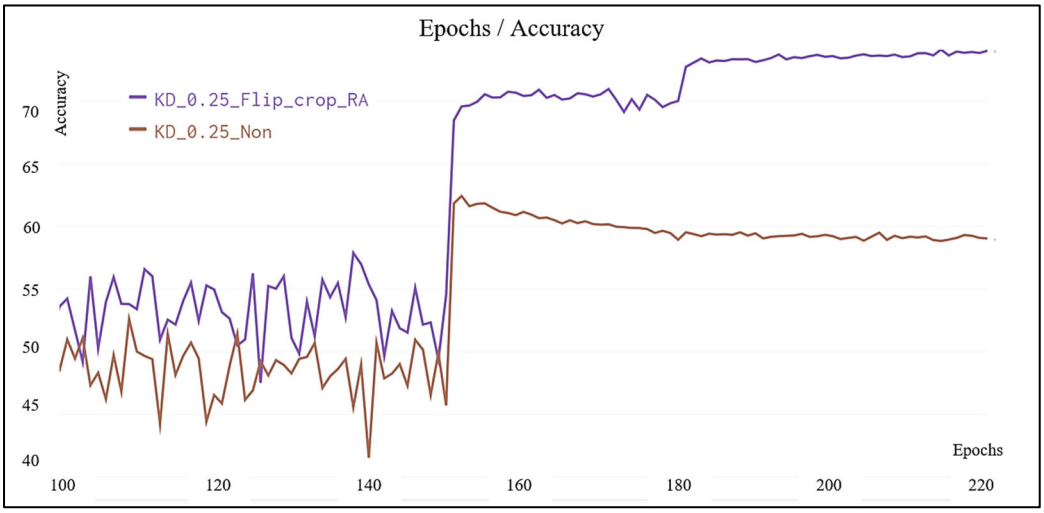

Figure 3. Difference between non(Brown) and Flip+crop+RA(Purple) when $alpha$ is 0.25

3 Conclusion

This paper conducted experiments on applying Data Augmentation to Knowledge Distillation (KD), a method for effectively training the Student Model, a relatively small architecture in the Image Classification field, and found that it significantly improved the model's performance.

We also tested the effect of different values of α on the performance of Data Augmentation. The results showed that the performance improvement of models trained with KD and Data Augmentation was substantial, and the impact of α value was minimal. Additionally, an interesting finding was observed that when the α value is 1.0, the performance improvement of Data Augmentation is relatively reduced compared to other α values.

References

Hinton, G., Vinyals, O. and Dean, J., *Distilling the knowledge in a neural network.* arXiv preprint arXiv:1503.02531, vol. 2, no. 7, 2015.

Cubuk, E. D., Zoph, B., Shlens, J. and Le, Q. V., Randaugment: Practical automated data augmentation with a reduced search space*. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 702-703, 2020.

Zagoruyko, S. and Komodakis, N*., Wide residual networks*. arXiv preprint arXiv:1605.07146, 2016.

Heo, B., Kim, J., Yun, S., Park, H., Kwak, N., and Choi, J. Y*., A comprehensive overhaul of feature distillation. In Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1921-1930, 2019.

Ridnik, T., Lawen, H., Ben-Baruch, E. and Noy, A., *Solving ImageNet: a Unified Scheme for Training any Backbone to Top Results*. arXiv preprint arXiv:2204.03475, 2022.

Fu, J., Geng, X., Duan, Z., Zhuang, B., Yuan, X., Trischler, A. and Dong, H*., Role-wise data augmentation for knowledge distillation*. arXiv preprint arXiv:2004.08861, 2020.

Yim, J., Joo, D., Bae, J. and Kim, J., *A gift from knowledge distillation*: Fast optimization, network minimization and transfer learning. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4133-4141, 2017.

Wang, H., Lohit, S., Jones, M. and Fu, Y. What Makes a "Good" *Data Augmentation in Knowledge Distillation–A Statistical Perspective,* 2019.

Gou, J., Yu, B., Maybank, S. J. and Tao, D., Knowledge distillation: A survey*. International Journal of Computer Vision*, vol. 129, no. 6, pp. 1789-1819, 2021.

Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C. and Bengio, Y., Fitnets*: Hints for thin deep nets. preprint arXiv*:1412.6550, 2014.

Cho, J. H. and Hariharan, B., *On the efficacy of knowledge distillation. In Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4794-4802, 2019.

Mirzadeh, S. I., Farajtabar, M., Li, A., Levine, N., Matsukawa, A. and Ghasemzadeh, H., *Improved knowledge distillation via teacher assistant. In Proceedings of the AAAI conference on artificial intelligence* , vol. 34, no. 04, pp. 5191-5198, 2020.

Beyer, L., Zhai, X., Royer, A., Markeeva, L., Anil, R. and Kolesnikov, A., Knowledge distillation: A good teacher is patient and consistent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* pp. 10925-10934, 2022.

Park, W., Kim, D., Lu, Y. and Cho, M., Relational knowledge distillation. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Feng, T. and Wang, M., *Response-based Distillation for Incremental Object Detection*. arXiv preprint arXiv:2110.13471, 2021.

Biography

Jun Seung Woo is student in MY PAUL SCHOOL. He is interested in artificial intelligence, deep learning, cryptography, block chains, autonomous vehicles, etc., and is conducting related research.

Shin Dong Ho is Professor and Teacher in MY PAUL SCHOOL. He obtained his Ph.D. in semiconductor physics in 2000. He is interested in artificial intelligence, deep learning, cryptography, robots, block chains, drones, autonomous vehicles, mechanical engineering, the Internet of Things, metaverse, virtual reality, and space science, and is conducting related research.