Attention Based Data Augmentation for Knowledge Distillation with Few Data

Shengzhao Tian¹, Duanbing Chen¹-², *  
¹Big Data Research Center, University of Electronic Science and Technology of China, Chengdu 611731, China  
²Chengdu Union Big Data Tech. Inc., Chengdu 610041, China  
*email: dbchen@uestc.edu.cn

Abstract. Knowledge distillation has attracted great attentions from computer vision researchers in recent years. However, the performance of student model will suffer from the absence of the complete dataset, which is used to train the teacher model. Especially for conducting knowledge distillation between heterogeneous models, it is difficult for student model to learn and receive guidance with few data. In this paper, a data augmentation method is proposed based on the attentional response of teacher model. The proposed method utilizes the knowledge in teacher model without requiring homogeneous architecture between teacher model and student model. Experimental results demonstrate that combining the proposed data augmentation method with different knowledge distillation methods, the performance of student model can be improved in knowledge distillation with few data.

1. Introduction

In recent years, the expensive calculation and training costs of deep models have raised concerns among researchers. Knowledge distillation is proposed to utilize the teacher model which has deep convolutional layers and massive parameters to guide the student model which is lighter and smaller in training process. The ultimate goal of knowledge distillation is making the light model solve problems with similar performance and lower computing costs than the heavy model.

Knowledge distillation methods can be divided into three categories according to the source of knowledge [1]: response-based, feature-based and relation-based. Specifically, response-based knowledge distillation [2] utilizes the output logits of the teacher model as the knowledge. The activations, neurons or features of intermediate layers are utilized as the knowledge to guide the learning of the student model in feature-based knowledge distillation [3-5]. Relation-based knowledge distillation [6-8] utilizes the relationships between different activations, neurons or pairs of samples as the knowledge.

In practical application scene, due to the objective causes such as unavailable of sensitive data, inconvenient data access and unexpected loss of data, the complete training data of teacher models may not be used during knowledge distillation. The absence of the complete dataset makes it difficult for student models to learn. Under this condition, the training effects of student models will become unsatisfactory. Naturally, this problem is gradually attracting the attention of researchers. Fang et al. [9] proposed a data-free adversarial distillation method inspired by the form of Generative Adversarial Networks (GANs). A good performance of image classification was achieved during distillation in which the teacher and student have the similar network structures. Nayak et al. [10] presented Zero-Shot Knowledge Distillation (ZSKD) in which the data impressions will be synthesized form the softmax space of teacher models and utilized as surrogates for the original training data samples. The...
student model could achieve a degree of classification performance without using any training data. Li et al. [11] presented few-sample knowledge distillation (FSKD), which is used in network compression where the student model is made by pruning the teacher model. Subsequently, Bai et al. [12] proposed a novel layer-wise knowledge distillation approach for effectively compressing network with few data. Recently, Shen et al. [13] proposed a novel grafting strategy for few-shot knowledge distillation. A dual-stage distillation scheme was used to replace the blocks of teacher model with the blocks of student model gradually and achieved gratifying results with few samples.

However, most of the previous works had achieved good results between teacher models and student models that have homogeneous architectures. Model compression methods such as pruning do not change the model structures. Conducting knowledge distillation between heterogeneous models with few data is still a challenge problem. In this paper, we explore knowledge distillation between heterogeneous models with few data from the perspective of data augmentation. A data augmentation method is proposed based on the attentional response of teacher model. The proposed method utilizes the effective response of teacher model to data. The knowledge containing in the teacher model will be used to enrich data. Experimental results demonstrate that our approach can improve the performance of student model in knowledge distillation with few data and does not require homogeneous architectures between teacher models and student models.

The main contributions of this paper are summarized as follows:

- A data augmentation method for knowledge distillation is proposed based on the attentional response of teacher model. Combining with several typical distillation methods, the performance breakthroughs on full dataset are achieved in knowledge distillation between heterogeneous models.
- The proposed data augmentation method can further improve the knowledge distillation performance on varying degrees of small datasets. Combining with the proposed method, the performances of several typical distillation methods have been improved when conducting knowledge distillation between heterogeneous models with few data.

The remains of the paper are organized as follows. The attentional response of teacher model, the proposed data augmentation method and the simplified cross-layer knowledge distillation method are described in Materials and Methods. The experimental results and analyses are shown in Results and Discussions. Conclusions and future works are given in Conclusions Section.

2. Materials and Methods

2.1. The attentional response of teacher model

Teacher models are usually trained well on full dataset. It means that the teacher model can perceive well on similar data. Such well perception is not only reflected in the output logits, but also in the activations of deep feature maps. Zagoruyko et al. [4] proved that the feature maps in the deep layers of a network can be mapped to the original image to express the attention of the network to images. Areas with high attentions will be used for classification by the model. Hence, the attentional response of model is defined as the mapping result of a deep feature map.

To generate the attentional response of a teacher model, firstly consider a deep layer and its activation tensor \( A \in \mathbb{R}^{C \times H \times W} \), where \( C \) represents feature channel and \( H \times W \) is the dimension of feature map. Then take the mean on channel dimension \( C \) to make a spatial attention map with dimension of \( H \times W \). Normalization will be carried out on the attention map. Finally, to project the spatial attention map to original image, the average pooling function is applied to generate the attentional response. A set of examples of attentional response are shown in figure 1. A well trained ResNet-32x4 [14] (79.42% top-1 accuracy) is used to generate attentional responses based on the activations of the last residual block. The key regions of the targets in images will generate high activation response. It demonstrates that the model has a strong ability to perceive data. We can use this capability to highlight key information in few training data and enrich data while keeping the key information remains.
Fig. 1. A set of examples of attentional response. Images are from CIFAR-100 dataset [15]. In each subgraph, the left one shows the original image, and the right one shows the original image with the attentional responses mask. The redder areas of the mask have higher activation values. Class of each subgraph are: (a) seal, (b) lion, (c) apple, (d) skunk.

2.2. Attention based data augmentation
The main idea of proposed data augmentation method is that the teacher model is utilized to sketch out the key information in limited data, and vary data while keeping the key information remains. Specifically, given a small training dataset, the size of each image is assumed to be $h \times w$. We run it on the teacher model so that the attentional responses of each image will be generated and saved. The attentional responses can be denoted as $AR \in \mathbb{R}^{h \times w}$. Then, the mask of attentional responses $MAR \in \mathbb{R}^{h \times w}$ can be expressed as:

$$MAR = \begin{cases} 0, & \text{if } AR < \text{mean}(AR) \\ 1, & \text{if } AR \geq \text{mean}(AR) \end{cases}$$

(1)

An element in $MAR$ is set to 0 when the element in the same position of $AR$ is less than the mean of $AR$, otherwise to 1. An augmented sample of an original image is defined as:

$$img_{augmented} = img_{original} \cdot MAR + img_{random} \cdot (One - MAR)$$

(2)

where $One \in \mathbb{R}^{h \times w}$ is a matrix where all element values is 1. $img_{random}$ is a random sample from the dataset. In summary, an augmented sample consists of key areas of the original image and random interference. The augmented samples increase the diversity of the dataset without losing target information. The problem that the model cannot learn enough generalization due to the small sample number is alleviated. An example of the augmented data is shown in figure 2.

2.3. Simplified cross-layer knowledge distillation
In this paper, we use a simplified cross-layer knowledge distillation modified from SemCKD [5]. Specifically, considering that it may become difficult to learn the attention allocation module with few samples, we remove the attention allocation module to reduce the parameters that need to be optimized. Each student layer that needs to be guided is directly associated with all target teacher layers through corresponding projection modules. Each projection module consists of a stack of three layers with 1x1, 3x3 and 1x1 convolutions to align the shapes of feature maps between a student layer and a target layer.
Fig. 2 An example of the attention based augmented data. Images are from CIFAR-100 dataset. Subgraph (a) is the original image. Subgraph (b) is the augmented image with random sample. Subgraph (c) is a random sample from the dataset.

3. Results and Discussions

In this paper, CIFAR100 [15] is used as experimental dataset which contains 100 classes. CIFAR100 has 60000 colour images with $32 \times 32$ size. 50000 images are used as training set and the rest 10000 images are used as test set. We randomly sample $k\%$ images per class in training set to construct small datasets, where $k \in \{10, 20, 30, 40, 50\}$. Following SemCKD [5], we adopt ResNet32x4 [14] as teacher model and VGG8 [16] as student model. The architectures of two models are significantly different. We consider them as a pair of heterogeneous models following Shen et al. [13].

Several typical one-stage heterogeneous distillation methods are used in experiments for comparison. No-KD represents training student model on the specified dataset without any knowledge distillation methods. VKD [2] represents the vanilla knowledge distillation which is a classic response-based distillation approach. SP [6] represents the similarity preserving knowledge distillation which is a typical relation-based distillation approach. SemCKD [5] is a state-of-the-art feature-based knowledge distillation approach. SCKD is the simplified cross-layer knowledge distillation modified from SemCKD [5]. ADA represents using the attention-based data augmentation proposed in this paper. None represents conducting knowledge distillation without any data augmentation. The experimental results of knowledge distillation with few data are shown in table 1.

Tab. 1 Experimental results of knowledge distillation with few data on CIFAR100 dataset. The teacher model is ResNet 32x4 [14] (Top-1 Accuracy: 79.42%), and the student model is VGG8 [16].

| KD Methods | DA Methods | $k$=10% | $k$=20% | $k$=30% | $k$=40% | $k$=50% | $k$=100% |
|------------|------------|---------|---------|---------|---------|---------|---------|
| None KD    | None       | 43.61%  | 53.81%  | 58.75%  | 61.66%  | 65.16%  | 71.18%  |
|            | ADA        | 43.70%  | 52.52%  | 58.93%  | 61.88%  | 64.36%  | 70.43%  |
| VKD [2]    | None       | 52.17%  | 60.17%  | 63.15%  | 65.15%  | 68.14%  | 72.90%  |
|            | ADA        | 55.76%  | 62.05%  | 66.03%  | 68.33%  | 70.40%  | 74.37%  |
| SP [6]     | None       | 53.86%  | 61.07%  | 64.56%  | 66.37%  | 68.95%  | 73.14%  |
|            | ADA        | 57.55%  | 64.43%  | 68.32%  | 69.63%  | 71.39%  | 74.49%  |
| SemCKD [5] | None       | 58.57%  | 64.29%  | 67.88%  | 69.62%  | 71.51%  | 75.27%  |
|            | ADA        | 61.82%  | 67.04%  | 70.84%  | 71.89%  | 72.82%  | 76.39%  |
| SCKD       | None       | 57.48%  | 64.13%  | 68.01%  | 69.98%  | 71.95%  | 76.01%  |
|            | ADA        | 62.88%  | 67.83%  | 70.82%  | 72.84%  | 73.88%  | 76.92%  |

Firstly, comparing whether knowledge distillation is used at different $k$, it can be found that using knowledge distillation always improves performance. The lower the $k$ value, the greater the boost knowledge distillation brings. For example, in $k=10\%$ and $k=50\%$, the top-1 accuracies of student models are 43.61% and 65.16% respectively without using KD and ADA. While using SCKD in the same $k$ value settings, the top-1 accuracies of student models increase by 13.87% and 6.79% respectively without using ADA. The benefits of KD in $k=10\%$ have almost doubled compared to
We believe that knowledge distillation itself has a certain ability to resist the absence of data, because the teacher model contains knowledge beyond the small dataset. Using knowledge distillation during training process can improve the performance of student model.

From experimental results, it can be seen that the attention-based data augmentation (ADA) proposed in this paper can further improve the knowledge distillation performance in small dataset. Using ADA with SCKD in $k=10\%$, the top-1 accuracy of student model increases by 5.4% compared to SCKD without ADA in the same $k$. Besides, we also find that the ADA can be compatible with different KD methods. The performances of three typical distillation methods in different types (response-based, feature-based and relation-based) are all improved with ADA in different $k$.

Combining the SCKD and ADA, a gratifying performance is achieved in knowledge distillation between heterogeneous models with few data. The highest performance is achieved for each $k$ value setting using SCKD with ADA compared to other experimental groups. Even under full dataset ($k=100\%$), the performance breakthroughs can prove the effectiveness of the proposed ADA method.

Besides, it can be seen from comparative experiments (None KD), ADA does not work well without KD. This phenomenon may indicate that the augmented samples can only play a role under the guidance of the teacher model. Student models may not make good use of these augmented samples by themselves.

In addition, to further understanding the mechanism of ADA, A comparative experiment is conducted with different kinds of random interferences. In ADA, the random interference comes from a randomly selected sample. An alternative is generating random values directly. We named the alternative strategy as ADA(RN), and named the original strategy as ADA(RS). Dataset and models remain unchanged and KD method is SCKD. The comparative experimental results of different kinds of random interferences are shown in table 2.

**Tab. 2** Comparative experimental results of different kinds of random interferences on CIFAR100 dataset. The teacher model is ResNet 32x4 [14], and the student model is VGG8 [16].

| KD Method | DA Methods | $k=10\%$ | $k=20\%$ | $k=30\%$ | $k=40\%$ | $k=50\%$ | $k=100\%$ |
|------------|------------|----------|----------|----------|----------|----------|----------|
| None       |            | 57.48\%  | 64.13\%  | 68.01\%  | 69.98\%  | 71.95\%  | 76.01\%  |
| SCKD       | ADA(RN)    | 59.01\%  | 65.90\%  | 70.24\%  | 71.33\%  | 73.34\%  | 76.37\%  |
|            | ADA(RS)    | **62.88\%** | **67.83\%** | **70.82\%** | **72.84\%** | **73.88\%** | **76.92\%** |

From experimental results, it can be seen that using random values as the random interference can still improve the performance of knowledge distillation in all $k$ value settings. Using ADA(RN) with SCKD in $k=30\%$, the top-1 accuracy of student model increases by 2.23% compared to using SCKD without ADA in the same $k$.

However, the performances of ADA(RN) are consistently inferior to the performances of ADA(RS) in all $k$ value settings. Especially in $k=10\%$, using ADA(RN) resulted in 3.87% decrease in the top-1 accuracy compared with using ADA(RS). The difference between random samples and random values is whether the interference has semantic information. Random sample has obvious semantic information and such semantic information makes model training challenging. Particularly when the amount of data is small, the model cannot learn enough generalization through limited samples. The confused augmented samples with semantic interferences can improve the generalization of the model effectively.

**4. Conclusion**

In this paper, we explore knowledge distillation between heterogeneous models with few data from the perspective of data augmentation. A data augmentation method is proposed based on the attentional response of teacher model. Combining the proposed data augmentation method with different knowledge distillation methods, gratifying performances are achieved in knowledge distillation between heterogeneous models with few data. Besides, we discuss some characteristics of the proposed method, including that this method needs to be used in conjunction with knowledge
distillation and the semantic random interferences is one explanation of how the proposed method improves the generalization of the model.

For future study, on the one hand, we will work on improving the performance of few-shot knowledge distillation between heterogeneous models with only one or five samples per class. On the other hand, we will intensively study the influence mechanism of knowledge in samples and teacher on student training process.

Acknowledgments
This research was funded by the National Natural Science Foundation of China with Grant No 61673085, the Fundamental Research for the Central Universities with Grant No ZYGX2019J074, Science Strength Promotion Programme of UESTC with Grant No Y03111023901014006, and the National Key Laboratory of Blind Signals Processing Innovation Development Fund with Grant No 61424132003.

References
[1] Gou J, Yu B, Maybank S J and Tao D 2021 Knowledge distillation: a survey Int. J. Comput. Vis. 129(6) pp 1789-1819
[2] Hinton G, Vinyals O and Dean J 2015 Distilling the knowledge in a neural network arXiv preprint arXiv: 1503.02531
[3] Romero A, Ballas N, Kahou S E, Chassang A, Gatta C and Bengio Y 2015 Fitnets: hints for thin deep nets Proc. Conf. Learn. Represent San Diego pp 1-13
[4] Zagoruyko S and Komodakis N 2017 Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer Proc. Int. Conf. Learn. Represent Toulon pp 1-13
[5] Chen D, Mei J P, Zhang Y, Wang C, Wang Z, Feng and Chen C 2021 Cross-layer distillation with semantic calibration Proc. Conf. AAAI Artif. Intell. 35(8) 7028-36
[6] Tung F and Mori G 2019 Similarity-preserving knowledge distillation Proc. IEEE Int. Conf. Comput. Vis. Seoul pp 1365-74
[7] Tian Y, Krishnan D and Isola P 2020 Contrastive representation distillation Proc. Int. Conf. Learn Represent pp 1-19
[8] Chen H, Wang Y, Xu C, and Tao D 2021 Learning student networks via feature embedding IEEE Trans. Neural Netw. Learn. Syst. 32(1) pp 25-35
[9] Fang G, Song J, Shen C, Wang X, Chen D and Song M 2019 Data-free adversarial distillation arXiv preprint arXiv 1912.11006
[10] Nayak G K, Mopuri K R, Shah V, Babu R V and Chakraborty A 2019 Zero-shot knowledge distillation in deep networks Proc. Int. Conf. Mach. Learn Long Beach pp 4743-51
[11] Li T, Li J, Liu Z and Zhang C 2020 Few sample knowledge distillation for efficient network compression Proc. IEEE Conf. Comput. Vis. Pattern Recog. pp 14639-47
[12] Bai H, Wu J, King I and Lyu M 2020 Few shot network compression via cross distillation Proc. Conf. AAAI Artif. Intell. New York 34(04) pp 3203-10
[13] Shen C, Wang X, Yin Y, Song J, Lou S and Song M 2021 Progressive network grafting for few-shot knowledge distillation Proc. Conf. AAAI Artif. Intell. 35(3) pp 2541-49
[14] He K, Zhang X, Ren S and Sun J 2016 Deep residual learning for image recognition Proc. IEEE Conf. Comput. Vis. Pattern Recog. Las Vegas pp 770-8
[15] Krizhevsky A and Hinton G 2009 Learning multiple layers of features from tiny images Technical Report
[16] Simonyan K and Zisserman A 2015 Very deep convolutional networks for large-scale image recognition Proc. Int. Conf. Learn. Represent San Diego pp 1-14