Improving Knowledge Distillation via Category Structure

Zailiang Chen[1], Xianxian Zheng[1], Hailan Shen[1,*], Ziyang Zeng[1], Yukun Zhou[2], and Rongchang Zhao[1]

[1] School of Computer Science and Engineering, Central South University, Changsha, Hunan, 410083, China
{xxxyczl,xxzheng,hailansh,zengziyang,zhaorc}@csu.edu.cn
[2] Centre for Medical Image Computing, University College London, London WC1V 6LJ, United Kingdom
yukun.zhou.19@ucl.ac.uk

Abstract. Most previous knowledge distillation frameworks train the student to mimic the teacher's output of each sample or transfer cross-sample relations from the teacher to the student. Nevertheless, they neglect the structured relations at a category level. In this paper, a novel Category Structure is proposed to transfer category-level structured relations for knowledge distillation. It models two structured relations, including intra-category structure and inter-category structure, which are intrinsic natures in relations between samples. Intra-category structure penalizes the structured relations in samples from the same category and inter-category structure focuses on cross-category relations at a category level. Transferring category structure from the teacher to the student supplements category-level structured relations for training a better student. Extensive experiments show that our method groups samples from the same category tighter in the embedding space and the superiority of our method in comparison with closely related works are validated in different datasets and models.

Keywords: knowledge distillation, intra-category structure, inter-category structure, structured relation

1 Introduction

Recent developments of deep neural network (DNN) have achieved state-of-the-art performance in many tasks [21, 1]. In several challenging datasets [11, 3], well-designed networks can even perform better than humans. However, these networks typically have millions of parameters and consume large amounts of computation resources. Applications of these large networks are limited on embedded devices due to their high resource demands. Therefore, there is an urgency for training small networks with low resource demands, while keeping the performance of small networks as close as possible to large networks. Several

*Corresponding Author.

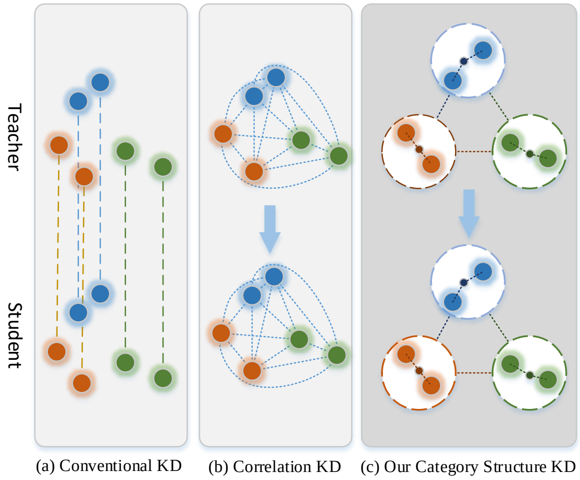


Fig. 1. Differences in transferred knowledge between conventional knowledge distillation, correlation knowledge distillation and our Category Structure Knowledge Distillation (CSKD). In contrast to previous methods, CSKD considers intra-category structure and inter-category structure at a category level as profitable knowledge for knowledge distillation to better improve the performance of the student

methods, such as low-rank factorization [10, 4], network pruning [15, 18], network quantization [13, 17] and knowledge distillation [8, 22], have been developed to solve this problem. Knowledge distillation has been proved to be an effective approach to improve the performance of small networks by transferring effective knowledge from a large model to a small model. Through additional regression constraints on outputs of teacher and student for input data, knowledge distillation forces the student model to imitate teacher's behaviors to obtain better performance.

The key problem of knowledge distillation is to extract effective, adequate and general knowledge from the teacher to the student. To handle this problem, conventional knowledge distillation transfers knowledge in a single-sample manner, keeping the student learning the consistency of each input sample as shown in Figure 1(a). It focuses on extracting knowledge from the final and immediate outputs of the teacher and transferring them to the student. Recently, correlation congruence [20] has been proposed to add constraints on relations between multiple samples as shown in Figure 1(b). However, these methods ignore the structured relations at a category level, which depict relations from a more abstract and high-level perspective.

We suppose that the category-level structured relations are also profitable knowledge for improving the performance of the student. In this paper, we further propose a novel general framework called Category Structure Knowledge Distillation (CSKD) which focuses on transferring category-level structured relations named category structure from the teacher to the student. Category structure consists of two types of structured relations: intra-category structure for each category and inter-category structure between different categories. Intra-category structure contains relations between samples from the same category and inter-category structure transfers relations between different categories. CSKD is easy to be implemented and the effectiveness of the proposed method is demonstrated by extensive empirical results on three datasets and different models.

Our contributions in this paper are summarized as follows:

1. We propose a new general distillation framework called Category Structure Knowledge Distillation (CSKD), which transfers structured relations from the teacher to the student at a category level. To the best of our knowledge, it is the first work to introduce category-level structured relations for knowledge distillation.
2. We define intra-category and inter-category structure to form the category structure. And two effective relation functions are introduced to better extract intra-category structure and inter-category structure from the embedding space of the teacher and the student.
3. Extensive experiments show that our method achieves state-of-the-art performance. We conduct experiments on different datasets and different teacher-student architecture settings to show the effectiveness of the proposed method in comparison with closely related works.

2 Related Work

In this paper, we focus on improving the performance of small networks. Therefore, we summarize recent methods in model compression and knowledge distillation in this section.

2.1 Model Compression

Model compression focuses on designing small networks with few parameters and high performance simultaneously. Sindhwani [24] proposed a unified framework to learn structured parameter matrices that are characterized by the notion of low displacement rank. Louizos [15] employed L_0 norm regularization in the training to prune the neural networks by encouraging weights to become exactly zero. Energy-aware pruning was utilized to construct energy-efficient convolutional neural networks [27]. Binary quantization with weights and activation constrained to {-1,+1} at run-time were adopted in [13]. Adaptive quantization for finding optimal quantization bit-width for each layer was also explored in recent work [30].

2.2 Knowledge Distillation

The purpose of knowledge distillation is improving the performance of small models by transferring knowledge from large models to small models. Hinton [8] first proposed to distill teacher's knowledge to student by soft targets under a controlled temperature. Romero [22] proposed a two-stage training procedure and transferred not only final outputs but also intermediate outputs to student. In [29], a compact student network was improved by mimicking the attention maps of a powerful teacher network. Yim [28] proposed a flow of solution procedure (FSP) to inherit relations between two convolutional layers. In [16], neurons in the deeper hidden layers were used to transfer essential characteristics of the learned face representation for face recognition task. To obtain a promising improvement, noise regularization was added while training the student [23]. Huang [9] regarded knowledge transfer as a distribution matching problem and utilized neuron selectivity patterns between teacher and student models to solve the distribution matching problem. In [7], activation boundary, which meant the activations of neurons instead of their exact output values, was employed to transfer classification-friendly partitions of the hidden feature space.

Recent works also adopt generative adversarial network (GAN) and adversarial examples to obtain better performance. In [26], conditional generative adversarial network was used to learn a proper loss function to transfer effective knowledge from teacher to student. And in [25], a three-player game, consisting of a teacher, a student, and a discriminator, was proposed based on generative adversarial network to force teacher and student learning each other mutually. Heo [6] forced student to learn the decision boundary by adversarial examples.

In addition to the above methods that transfer knowledge in a single-sample manner, there are also a few methods to explore relations between multiple samples for knowledge distillation. Chen [2] used cross-sample similarities which could be naturally derived from deep metric models. In [19], distance-based and angle-based relations were proposed to penalize structural differences in relations. Similarly, Liu [14] utilized instance relationship graph to transfer a relation graph from teacher to student.

In this paper, we further take the category structure in feature space as profitable knowledge to transfer the intra-category structure and inter-category structure from teacher to student.

3 Category Structure Knowledge Distillation

In this section, we describe the details of our proposed category structure for knowledge distillation.

3.1 Knowledge Distillation

We start from conventional knowledge distillation in this section for a better understanding. The concept of knowledge distillation is first proposed in [8] to distill hint knowledge from teacher to the student using cross-entropy

$$\mathcal{L}_{KD-CE} = \frac{1}{n}\sum_{i=1}^{n}\mathcal{H}_{cross}(\boldsymbol{y_i^t}, \boldsymbol{y_i^s}), \tag{1}$$

where n is the number of samples and $\mathcal{H}_{cross}$ is cross-entropy loss function. $\boldsymbol{y_i^t}$ and $\boldsymbol{y_i^s}$ refer to teacher's and student's softmax outputs under distillation temperature τ

$$y_{ij} = \frac{e^{z_j/\tau}}{\sum_{k=1}^{c} e^{z_k/\tau}}, \tag{2}$$

where y_{ij} refers to the predicted probability belonging to the j-th class, z_j refers to the logits of teacher and student and c represents the number of classes. By minimizing the cross-entropy loss function, student mimics teacher's behaviors progressively. In several works [20, 7], KL divergence is adopted to better match distributions of teacher and student.

$$\mathcal{L}_{KD-KL} = \frac{1}{n}\sum_{i=1}^{n} \mathcal{KL}(\boldsymbol{y_i^t}, \boldsymbol{y_i^s}). \tag{3}$$

Correlation constraints are utilized in [20] to transfer relations between multiple samples by computing cross sample correlations.

$$\mathcal{L}_{correlation} = \frac{1}{n^2}\|\Phi(\boldsymbol{F^t}) - \Phi(\boldsymbol{F^s})\|_2^2, \tag{4}$$

where $\boldsymbol{F^t}$ and $\boldsymbol{F^s}$ represent feature maps of teacher and student, respectively. $\Phi(\cdot)$ is a mapping function, $\Phi : \boldsymbol{F} \rightarrow \boldsymbol{\Omega} \in \mathbb{R}^{n\times n}$, which maps feature representation $\boldsymbol{F}$ to a relational matrix $\boldsymbol{\Omega}$ by computing pairwise similarity or distance between any two samples in a mini-batch of training dataset. Correlation reflects relations between samples and transferring mutual correlation to student can improve the performance of student by providing extra beneficial information that can not be noticed in single sample manner.

Transferring pairwise relations between any two samples is straight-forward and it may contain some redundant and irrelevant information for knowledge distillation. For example, relations between samples from different classes are calculated for any pair in [20]. Samples from highly related classes may get high similarity and samples from irrelevant classes may get low similarity. However, most of these relations are redundant and unnecessary for classification task. Samples from the same class may have similar relations between themselves and samples from other classes. Transferring redundant information from teacher to student may confuse student to some extent. Inspired by this, we consider structured relations at a category level as principal and sparse knowledge. Beyond sample correlation, we further explore category structure for knowledge distillation.

3.2 Category Structure

In this section, we describe Category Structure Knowledge Distillation in detail. Category structure consists of two parts: intra-category structure and inter-category structure. Intra-category structure describes structured relations between samples from the same category, while inter-category structure represents

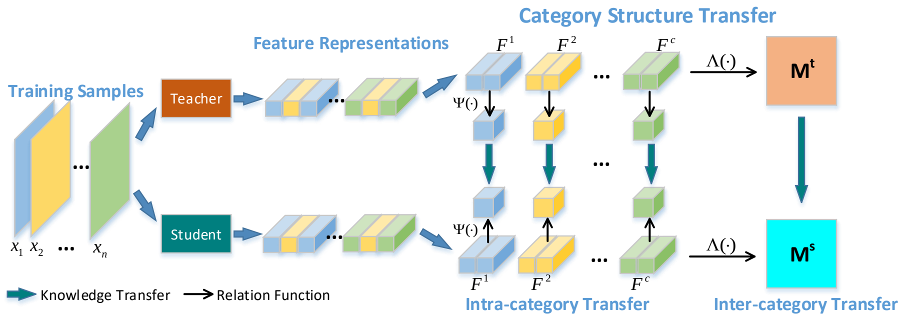


Fig. 2. The overview of our CSKD. Extract intra-category structure and inter-category structure by relation functions $\Psi(\cdot)$ and $\Lambda(\cdot)$ respectively, and transfer them from the teacher to the student

structured relations between different categories at a category level. The overall framework of our proposed method is illustrated in Figure 2. Given n training samples $\boldsymbol{X} = \{x_1, x_2, ..., x_n\}$, a pre-trained teacher model f^t and a random initialized student model f^s. Let feature representations $\boldsymbol{F}^{it} = f^t(\boldsymbol{X}^i; \boldsymbol{W}^t)$ and $\boldsymbol{F}^{is} = f^s(\boldsymbol{X}^i; \boldsymbol{W}^s)$, respectively. $\boldsymbol{W}^t$ and $\boldsymbol{W}^s$ are weights of teacher and student. $\boldsymbol{X}^i$ refers to samples belonging to the i-th class. We divide training samples into different categories by labels. Then category structure denoted as CS is constructed to represent relation structures across samples and can be expressed as

$$CS = (CS_{intra}, CS_{inter}) = (\{\Psi(\boldsymbol{F}^i)\}_{i=1}^c, \Lambda(\{\boldsymbol{F}^i\}_{i=1}^c)), \tag{5}$$

where $\Psi(\cdot)$ is the intra-category structure function constructing relations between samples from the same category and $\Lambda(\cdot)$ refers to the inter-category structure function representing relations between different categories. For each feature representation set $\boldsymbol{F}^i = f(\boldsymbol{X}^i; \boldsymbol{W})$ belonging to the i-th category, $\Psi(\boldsymbol{F}^i)$ formalise their structured relations to group a tight cluster in the embedding space. Correspondingly, $\Lambda(\{\boldsymbol{F}^i\}_{i=1}^c))$ is a mapping function: $\Lambda : \boldsymbol{F} \rightarrow \boldsymbol{M} \in \mathbb{R}^{c \times c}$, calculating similarities between different categories to separate samples from irrelevant categories from each other. $\boldsymbol{M}$ is a category relational matrix.

To construct relations at a category level, we define a category center as

$$\boldsymbol{C}^i = \frac{1}{m} \sum_{j=1}^{m} \boldsymbol{F}_j^i, \tag{6}$$

where $\boldsymbol{F}_j^i$ refers to the feature map belonging to the j-th sample from the i-th category, and m is the number of samples from the i-th category. Category center is calculated by the average feature map for samples from the same category and it represents the general category feature representation in high-level feature space to some extent. Then the relation function of intra-category structure can

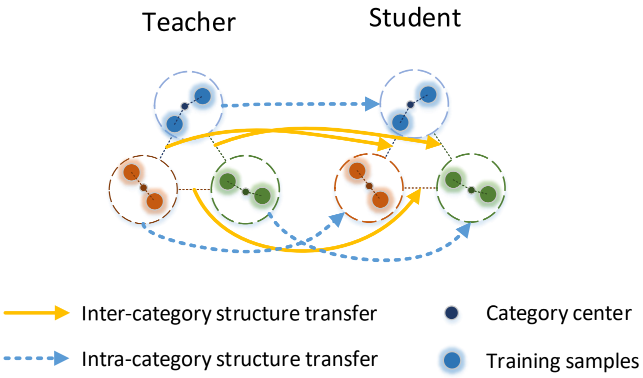

Teacher
Student
Inter-category structure transfer
Intra-category structure transfer
Category center
Training samples

Fig. 3. Illustration of category structure transfer. Yellow solid arrows indicate the inter-category structure transfer, and blue dotted arrows refer to intra-category structure transfer. Inter-category structure transfers cross-category similarity between any two categories and intra-category structure transfer relative structure formed by category center and training samples from the same category

be defined as

$$\Psi(\boldsymbol{F}^i) = \{\boldsymbol{F}^i_j - \boldsymbol{C}^i\}^m_{j=1}. \tag{7}$$

It preserves the structured information of relative distances between each sample and its category center. We assume that samples from the same category group tight in the embedding space and category center can represent samples from the same category in the embedding space. Based on category center, relations of samples from the same category are involved in the relation structure in a more efficient and sparse way. We further define the relation function of inter-category structure based on similarity:

$$\boldsymbol{M}(i,j) = \Lambda(\boldsymbol{F}^i, \boldsymbol{F}^j) = \frac{\boldsymbol{C}^i \cdot \boldsymbol{C}^j}{\|\boldsymbol{C}^i\|_2 \|\boldsymbol{C}^j\|_2}, i,j = 1,2,...,c. \tag{8}$$

It reflects the structured relations between any two categories. Highly related categories have high similarity scores and irrelevant categories have low similarity scores.

Intra-category structure ignores redundant relations between cross-category samples and focuses on pairwise relations formed by relative distance to category center between samples from the same category. Correspondingly, inter-category structure maintains principal category-wise relations and it complements structured relations in a global sense. And our extensive experiments shows that intra-category structure and inter-category structure shows mutual positive effects to each other. Since we conduct structured relations at a category level, category structure constructs sparser relations than correlation that calculates relations between any two samples (see analysis in Section 4.6).

3.3 Loss for Category Structure Transfer

Figure 3 shows the illustration of category structure and its transfer process.To transfer category structure from teacher to student, we construct L_{CS} to measure differences between category structures of teacher and student. Let $D(\cdot)$ represents the distance function between relation structures, then the loss for category structure transfer can be defined as

$$\begin{aligned} L_{CS} &= L_{intra} + L_{inter} \\ &= \beta \cdot D(CS^t_{intra}, CS^s_{intra}) + \gamma \cdot D(CS^t_{inter}, CS^s_{inter}) \\ &= \frac{\beta}{c} \cdot \sum_{i=1}^{c} \|\varPsi(\boldsymbol{F}^{it}) - \varPsi(\sigma(\boldsymbol{F}^{is}))\|_2 + \gamma \cdot \|\varLambda(\{\boldsymbol{F}^{it}\}_{i=1}^{c}) - \varLambda(\{\boldsymbol{F}^{is}\}_{i=1}^{c})\|_2 \quad (9) \\ &= \frac{\beta}{c} \cdot \sum_{i=1}^{c} \|\varPsi(\boldsymbol{F}^{it}) - \varPsi(\sigma(\boldsymbol{F}^{is}))\|_2 + \gamma \cdot \|\boldsymbol{M}^t - \boldsymbol{M}^s\|_2, \end{aligned}$$

where β and γ are hyper-parameters to control weights of intra-category structure and inter-category structure. $\sigma(\cdot)$ is a transformer with 1×1 convolution layer for matching student's channels to teacher's. Therefore, total loss for training student is

$$L_{total} = \alpha L_{CE} + (1 - \alpha) L_{KD} + L_{CS}, \quad (10)$$

where L_{CE} is the cross-entropy loss based on student's output and ground truth and L_{KD} is the mean square errors of teacher's and student's logits in our experiments. α is a trade-off between supervision from labels and single-sample based knowledge transfer. Our CSKD is summarized in Algorithm 1.

4 Experiments

We evaluate CSKD on three datasets: CIFAR-10, CIFAR-100 and Tiny ImageNet to show the effectiveness of our proposed method. And we compare CSKD with closely related works. Extensive experiments are conducted to explore category structure for knowledge distillation. Our codes for experiments and more results will be available at https://github.com/xeanzheng/CSKD.

Algorithm 1 Category structure knowledge distillation.

```
Input:
    Training samples, $\boldsymbol{X} = \{x_1, x_2, ..., x_n\}$;
    Labels of training samples, $\boldsymbol{y} = \{y_1, y_2, ..., y_n\}$;
    Teacher model $f^t$ with pre-trained weights $\boldsymbol{W^t}$;
    Student model $\hat{f}^s$ with random initialized weights $\boldsymbol{\hat{W}^s}$;
    Transformer with 1×1 convolution layer, $\sigma$;
Output:
    Student model $f^s$ with optimized weights $\boldsymbol{W^s}$;
while not convergence do
    Choose a random batch $\hat{\boldsymbol{X}}$ and their labels $\hat{\boldsymbol{y}}$ from training samples $\boldsymbol{X}$ and labels $\boldsymbol{y}$;
    Extract features from teacher and student model, $\boldsymbol{F}^t = f^t(\hat{\boldsymbol{X}}; \boldsymbol{W}^t)$, $\boldsymbol{F}^s = \hat{f}^s(\hat{\boldsymbol{X}}; \boldsymbol{W}^s)$;
    Group features into different groups by labels $\hat{\boldsymbol{y}}$, $\boldsymbol{F}^t = \{\boldsymbol{F}^{it}\}_{i=1}^c$, $\boldsymbol{F}^s = \{\boldsymbol{F}^{is}\}_{i=1}^c$;
    Extract structured relations by $\Psi$ and $\Lambda$ relation functions, $\Psi(\boldsymbol{F}^{it})$, $\Psi(\sigma(\boldsymbol{F}^{is}))$, $\Lambda(\{\boldsymbol{F}^{it}\}_{i=1}^c)$, $\Lambda(\{\boldsymbol{F}^{is}\}_{i=1}^c)$;
    Transfer category structure to student model $f^s$ by descending the stochastic gradient from $L_{CS}$:
        $\nabla_{\boldsymbol{W}^s} \frac{\beta}{c} \cdot \sum_{i=1}^c `\|\Psi(\boldsymbol{F}^{it}) - \Psi(\sigma(\boldsymbol{F}^{is}))\|_2 + \gamma \cdot \|\Lambda(\{\boldsymbol{F}^{it}\}_{i=1}^c) - \Lambda(\{\boldsymbol{F}^{is}\}_{i=1}^c)\|_2$,
    and train the student model by supervision from labels and single-sample based knowledge distillation loss from the teacher:
        $\nabla_{\boldsymbol{W}^s} \alpha L_{CE} + (1 - \alpha) L_{KD}$;
end while
return $f^s$ with its weights $\boldsymbol{W}^s$;
```

4.1 Experimental Settings

We adopt ResNet [5] as the main architecture in our experiments. In the main experiments, the hyper-parameter α is set to 0.1, the weight of intra-category structure loss β is empirically set to 0.01, and $\gamma = 0.2$.

On CIFAR-10, CIFAR-100, and Tiny ImageNet, we compare CSKD with the student trained with only cross-entropy (CE), original knowledge distillation (KD) [8], Fitnet [22], KDGAN [25], activation boundary transfer (AB) [7], and correlation congruence knowledge distillation (CCKD) [20]. For fare comparisons, all methods are implemented and compared under the same architecture configurations.

4.2 Results on CIFAR-10

CIFAR-10 [12] consists of 60K 32×32 images in 10 classes and each class contains 5000 images in training set and 1000 images in validation set. We first resize images to 40×40 by zero-padding, and then randomly crop images to original size 32×32. Meanwhile, random horizontal flip and normalization with channel means and standard deviations are adopted to augment the training data. We use a batch size 128 and a standard SGD optimizer with an initial learning rate 0.1 and momentum 0.9 to optimize our model and the weight decay is set to

Table 1. Accuracy of different methods on CIFAR-10. We explore our CSKD on different teacher-student architecture settings and keep the same training configuration for all the methods for fair comparisons. The proposed method surpasses all other methods. R101_0.5: ResNet101 with a channel reduction to the ratio of 50%

Teacher/Student Model	CE	KD	Fitnet	KDGAN	AB	CCKD	Proposed	Teacher
R101_0.5/R18_0.25	90.14	91.01	91.05	91.54	91.42	91.07	**91.99**	92.91
R101_0.5/R34_0.25	91.25	91.48	91.61	92.09	91.94	91.87	**92.67**	92.91
R101_0.5/R50_0.25	92.16	92.18	92.37	92.65	92.49	92.34	**93.20**	92.91
R152_0.5/R18_0.25	90.14	91.11	91.37	91.50	91.44	91.56	**92.27**	93.22
R152_0.5/R34_0.25	91.25	91.81	92.14	92.38	92.16	92.10	**92.76**	93.22
R152_0.5/R50_0.25	92.16	92.29	92.53	93.01	92.79	92.75	**93.31**	93.22

1e-4. We train the model with 200 epochs and the learning rate is multiplied by a scale factor 0.1 when training epochs are at 80, 120, 160.

We conduct our CSKD on teacher networks ResNet152_0.5 (14.6M) with a accuracy 93.22% and ResNet101_0.5 (10.7M) with a accuracy 92.91% and student networks ResNet18_0.25 (0.7M), ResNet34_0.25 (1.3M) and ResNet50_0.25 (1.4M). ResNet_x represents a ResNet with a channel reduction to a ratio of x. The first convolution kernel is changed to size 3×3 with a stride 1 and the stride of the first max-pooling is set to 1 to fit the image size.

We show our results on CIFAR-10 in Table 1. CSKD shows remarkable improvements under all evaluated teacher-student architecture settings. It obtains an average 1.52% improvement on different student networks and surpasses several closely related state-of-the-art methods with obvious margins. And it is noticed that our compression ratios are around 4.8%∼13.1%, however, the performance of the student even can surpass the teacher in some teacher-student architecture settings, e.g., 92.91% of teacher ResNet101_0.5 versus 93.20% of student ResNet50_0.25.

Table 2. Accuracy of different methods on CIFAR-100. Our CSKD outperforms all other methods and even better than the teacher

Teacher/Student Model	CE	KD	Fitnet	KDGAN	AB	CCKD	Proposed	Teacher
R101_0.5/R18_0.25	65.64	67.43	68.04	68.35	68.17	68.96	**69.14**	71.77
R101_0.5/R34_0.25	66.86	69.30	69.76	69.81	69.91	70.14	**70.39**	71.77
R101_0.5/R50_0.25	68.79	70.57	71.39	71.24	71.05	71.32	**71.61**	71.77
R152_0.5/R18_0.25	65.64	67.99	68.41	68.34	68.73	69.15	**69.22**	72.15
R152_0.5/R34_0.25	66.86	70.08	70.48	70.70	70.75	70.98	**71.01**	72.15
R152_0.5/R50_0.25	68.79	71.24	71.92	71.52	72.25	72.19	**72.60**	72.15

Table 3. Top-1 accuracy and top-5 accuracy on Tiny ImageNet. The teacher is ResNet152 (58.5M) and the student is ResNet18_0.25 (0.7M)

Method	Top-1 accuracy	Top-5 accuracy
Teacher	60.70	81.87
CE	45.21	71.03
KD	49.53	74.90
Fitnet	50.12	75.41
KDGAN	52.84	77.62
CCKD	53.14	78.14
AB	52.72	77.89
Proposed	**53.66**	**78.75**

4.3 Results on CIFAR-100

CIFAR-100 [12] is similar to CIFAR-10 dataset. But it is a more complicated dataset because it contains 100 classes rather than 10 classes in CIFAR-10. There are also 60K 32×32 images in CIFAR-100 and 50K/10K images for training/validation. Each class contains 500 training images and 100 validation images and we adopt the same data augmentation scheme used in CIFAR-10 (resize/padding/crop/flip/normalization) for CIFAR-100. The same multi-step SGD optimizer is also adopted and we train our model with 200 epochs.

We show results on CIFAR-100 in Table 2. The same network architecture settings are used and CSKD outperforms other methods. In this dataset, there exists a relatively big margin (compression ratio around 2.98%∼6.51%) between teacher and student and our CSKD improves the performance of the student by 2.82%∼4.15%. And the student achieves a better accuracy 72.60% (ResNet50_0.25) than the teacher with an accuracy 72.15% (ResNet152_0.5) at the last entry in Table 2.

4.4 Results on Tiny ImageNet

Tiny ImageNet is a downsampled version of the ImageNet [3] for classification. It consists of 120K images with 200 classes and each class contains 500 training images, 50 validating images, and 50 test images. The images are downsampled from 256×256 to 64×64. It is more difficult to classify these images in Tiny ImageNet than CIFAR datasets. We adopt Resnet152 (58.5M) as the teacher model and Resnet18_0.25 (0.7M) as the student model to explore the performance of CSKD when there is a big gap in capacity between the teacher and the student. The student only has around 1.2% of the teacher's parameters under this teacher-student architecture setting. We resize input images to 72×72 and then randomly crop them to 64×64. Random horizontal flip operation and channel normalization are also utilized to augment and normalize the training data. To better extract feature representations in the embedding space, the first convolutional kernel in original ResNet18 is changed to 3×3 with a stride 1 to fit the image size. The batch size is chosen as 200 and the student is trained with 200 epochs. A SGD optimizer with initial learning rate 0.1 and momentum 0.9 is utilized and the weight decay is set to 5e-4. The learning rate is divided by a factor 10 at 50, 100, 150 epochs.

Table 3 shows the results of CSKD and related works on Tiny ImageNet. All models are evaluated on validation set and trained with the same epochs for fair comparisons. Our CSKD surpasses all other methods in Table 3 and gets a 53.66% top-1 accuracy and a 78.75% top-5 accuracy. Compared with original KD, CSKD surpasses by around 4% in both top-1 accuracy and top-5 accuracy.

Table 4. Ablation study of Category Structure Knowledge Distillation. It is observed that every part of our category structure takes effect. Intra-category structure and inter-category structure show mutual effects when both of them are used

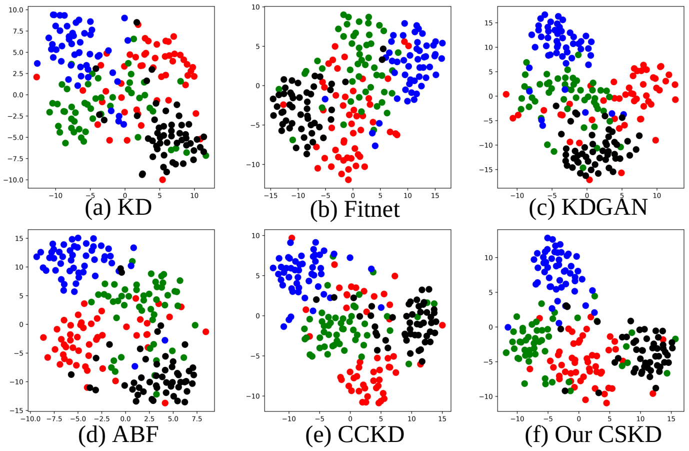
Intra loss	Inter loss	Top-1 accuracy	Top-5 accuracy
×	×	49.53	74.90
✓	×	52.51	78.36
×	✓	52.48	78.45
✓	✓	**53.66**	**78.75**

4.5 Ablation Study

We conduct an ablation study on the setting of a teacher ResNet152 and a student ResNet18_0.25 to delve into two parts of category structure, i.e., intra-category structure and inter-category structure. The results are summarized in Table 4. Each part of category structure loss is stripped to show the effectiveness of two parts of our category structure. When applied only intra-category loss or inter-category loss, our method gets similar improvements. If unabridged category structure loss is used, intra-category loss and inter-category loss show mutual effects on each other and CSKD achieves better promotions. It is also noticed that our method gets a general higher top-5 accuracy when compared with all other methods in Table 3, which reveals that category structure groups similar categories tighter in the embedding space and separates irrelevant categories far away from each other.

4.6 Analysis

Since we construct relation structures at a category level, the relations are sparser than cross-sample correlation which penalizes relations between any two samples. We simply regard different kinds of relations as the same edges between different vertices (samples) and calculate the number of edges to compare the complexity between category structure and cross-sample correlation. Let a dataset consists



(a) KD (b) Fitnet (c) KDGAN

(d) ABF (e) CCKD (f) Our CSKD

Fig. 4. Feature visualization of different methods. Each point represents a sample and each color represents a class. Our CSKD groups samples from the same category tighter than all other methods. Best viewed in color

of c categories and to simplify the calculation, each category is assumed to contain m images, then the number of edges in correlation is m^2c^2. In category structure, the number of edges is $mc + c^2$. Then the quantity ratio can be calculated as

$$m^2c^2/(mc + c^2) = m^2c/(m + c) \leq \frac{m}{2}\sqrt{mc}. \tag{11}$$

It is obvious that category structure compresses original correlation at most $\frac{m}{2}\sqrt{mc}$ times. Our CSKD helps reduce redundant relations between cross-category samples by focusing on relations between samples from the same category (intra-category relations) and using relations based on category center between different categories (inter-category relations).

To better show the effect of our CSKD, we extract feature representations of the last layer in student ResNet18_0.25 and visualize them as shown in Figure 4. A random batch of validation set in Tiny ImageNet is used and therefore, there are only four random classes for a clear comparison. It is observed that CSKD group samples from the same category tighter than all other methods (e.g., the green clusters in Figure 4) and each cluster in CSKD has a relatively clear boundary to other clusters.

5 Conclusion

In this paper, we find that relation transfer for knowledge distillation can be further explored at a category level. For classification tasks, the concept of category can be easily defined by labels. So we construct intra-category structure and inter-category structure based on labels to transfer principal relational knowledge in a sparse but powerful way. Intra-category structure preserves the structured relations in samples from the same category, while inter-category structure reflects the cross-category relations at a category level.

Our CSKD is implemented in a mini-batch, which may be a limitation when the number of categories is close to batch size. In this case, our category structure may degrade to the cross-sample correlation that transfers relations between any two samples. We set batch size equal to or larger than the number of classes to ensure that our CSKD takes effect in our experiments. And for the sake of fair comparisons, we set the batch sizes of all other methods to the same as our CSKD. An alternative to address this issue is to construct a better sampler for training and we will explore this problem in future. Another issue worth exploring is that our CSKD naturally fits the setting of multi-label classification tasks because of the existence of more complex and strong cross-category relations in multi-label classification datasets. Implementing CSKD in multi-label classification tasks may get more convincing improvement.

Acknowledgements. This work is supported by the National Natural Science Foundation of China (61972419, 61702558, 61672542), Natural Science Foundation of Hunan Province of China (2020JJ4120), and Fundamental Research Funds for the Central Universities of Central South University (2019zzts963).

References

1. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence **40**(4), 834–848 (2017)
2. Chen, Y., Wang, N., Zhang, Z.: Darkrank: Accelerating deep metric learning via cross sample similarities transfer. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. IEEE (2009)
4. Denton, E.L., Zaremba, W., Bruna, J., LeCun, Y., Fergus, R.: Exploiting linear structure within convolutional networks for efficient evaluation. In: Advances in neural information processing systems. pp. 1269–1277 (2014)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
6. Heo, B., Lee, M., Yun, S., Choi, J.Y.: Knowledge distillation with adversarial samples supporting decision boundary. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 3771–3778 (2019)
7. Heo, B., Lee, M., Yun, S., Choi, J.Y.: Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 3779–3787 (2019)

8. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
9. Huang, Z., Wang, N.: Like what you like: Knowledge distill via neuron selectivity transfer. arXiv preprint arXiv:1707.01219 (2017)
10. Jaderberg, M., Vedaldi, A., Zisserman, A.: Speeding up convolutional neural networks with low rank expansions. arXiv preprint arXiv:1405.3866 (2014)
11. Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R.: Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2901–2910 (2017)
12. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images. Tech. rep., Citeseer (2009)
13. Lin, X., Zhao, C., Pan, W.: Towards accurate binary convolutional neural network. In: Advances in Neural Information Processing Systems. pp. 345–353 (2017)
14. Liu, Y., Cao, J., Li, B., Yuan, C., Hu, W., Li, Y., Duan, Y.: Knowledge distillation via instance relationship graph. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7096–7104 (2019)
15. Louizos, C., Welling, M., Kingma, D.P.: Learning sparse neural networks through l_0 regularization. arXiv preprint arXiv:1712.01312 (2017)
16. Luo, P., Zhu, Z., Liu, Z., Wang, X., Tang, X.: Face model compression by distilling knowledge from neurons. In: Thirtieth AAAI Conference on Artificial Intelligence (2016)
17. Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., et al.: Mixed precision training. arXiv preprint arXiv:1710.03740 (2017)
18. Molchanov, P., Tyree, S., Karras, T., Aila, T., Kautz, J.: Pruning convolutional neural networks for resource efficient inference. arXiv preprint arXiv:1611.06440 (2016)
19. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3967–3976 (2019)
20. Peng, B., Jin, X., Liu, J., Li, D., Wu, Y., Liu, Y., Zhou, S., Zhang, Z.: Correlation congruence for knowledge distillation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5007–5016 (2019)
21. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
22. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550 (2014)
23. Sau, B.B., Balasubramanian, V.N.: Deep model compression: Distilling knowledge from noisy teachers. arXiv preprint arXiv:1610.09650 (2016)
24. Sindhwani, V., Sainath, T., Kumar, S.: Structured transforms for small-footprint deep learning. In: Advances in Neural Information Processing Systems. pp. 3088–3096 (2015)
25. Wang, X., Zhang, R., Sun, Y., Qi, J.: Kdgan: knowledge distillation with generative adversarial networks. In: Advances in Neural Information Processing Systems. pp. 775–786 (2018)
26. Xu, Z., Hsu, Y.C., Huang, J.: Training shallow and thin networks for acceleration via knowledge distillation with conditional adversarial networks. arXiv preprint arXiv:1709.00513 (2017)

27. Yang, T.J., Chen, Y.H., Sze, V.: Designing energy-efficient convolutional neural networks using energy-aware pruning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5687–5695 (2017)
28. Yim, J., Joo, D., Bae, J., 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)
29. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv preprint arXiv:1612.03928 (2016)
30. Zhou, Y., Moosavi-Dezfooli, S.M., Cheung, N.M., Frossard, P.: Adaptive quantization for deep neural network. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)