Few-Shot Image Recognition with Knowledge Transfer

Zhimao Peng[†], Zechao Li[†]*, Junge Zhang[‡], Yan Li[‡], Guo-Jun Qi[♯], Jinhui Tang[†]
[†]School of Computer Science and Engineering, Nanjing University of Science and Technology
[‡]Institute of Automation, Chinese Academy of Sciences [♯]Huawei Cloud
{zhimaopeng, guojunq}@gmail.com, {zechao.li, jinhuitang}@njust.edu.cn
jgzhang@nlpr.ia.ac.cn, yan.li@cripac.ia.ac.cn

Abstract

Human can well recognize images of novel categories just after browsing few examples of these categories. One possible reason is that they have some external discriminative visual information about these categories from their prior knowledge. Inspired from this, we propose a novel Knowledge Transfer Network architecture (KTN) for few-shot image recognition. The proposed KTN model jointly incorporates visual feature learning, knowledge inferring and classifier learning into one unified framework for their optimal compatibility. First, the visual classifiers for novel categories are learned based on the convolutional neural network with the cosine similarity optimization. To fully explore the prior knowledge, a semantic-visual mapping network is then developed to conduct knowledge inference, which enables to infer the classifiers for novel categories from base categories. Finally, we design an adaptive fusion scheme to infer the desired classifiers by effectively integrating the above knowledge and visual information. Extensive experiments are conducted on two widely-used Mini-ImageNet and ImageNet Few-Shot benchmarks to evaluate the effectiveness of the proposed method. The results compared with the state-of-the-art approaches show the encouraging performance of the proposed method, especially on 1-shot and 2-shot tasks.

1. Introduction

Recently classical deep learning models have achieved remarkable success on many computer vision and image understanding tasks [11, 7, 16, 27, 15]. To further improve the performance, the neural networks become deeper, which usually requires more labeled data in richer categories. Unfortunately, not only the human-labeled data is often very expensive, but the classical deep learning methods easily have the problems of overfitting and poor generalization capability with limited labeled data.

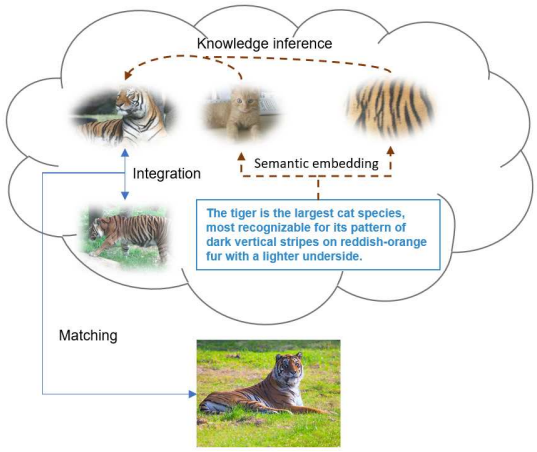

Figure 1. Given a labeled tiger image, people have some prior knowledge about tiger (such as "cat specie" and "dark vertical stripes") and then conduct knowledge inference to generate the discriminative visual information for tiger. The proposed method imitates this process to improve few-shot recognition performance.

To alleviate the demand of labeled data for deep model training, few-shot learning has attracted wide attention recently [33, 3, 24, 29, 39, 18, 36, 6, 23, 5]. The goal of few-shot learning is to recognize novel categories with only one or few labeled examples. The key idea is to transfer visual patterns obtained from base categories to describe the novel categories. Most existing few-shot learning methods combined with deep learning can be roughly divide into two groups: metric-learning based methods [10, 33, 29, 39] and meta-learning based methods [3, 24, 18, 36, 39]. Metric-learning based methods mainly focus on learning an appropriate visual feature embedding space with deep networks and choosing a well-defined metric to calculate the similarities between the few examples of novel categories and testing examples. Meta-learning based methods try to learn some transferable "meta knowledge" from past experiences

*Corresponding author.

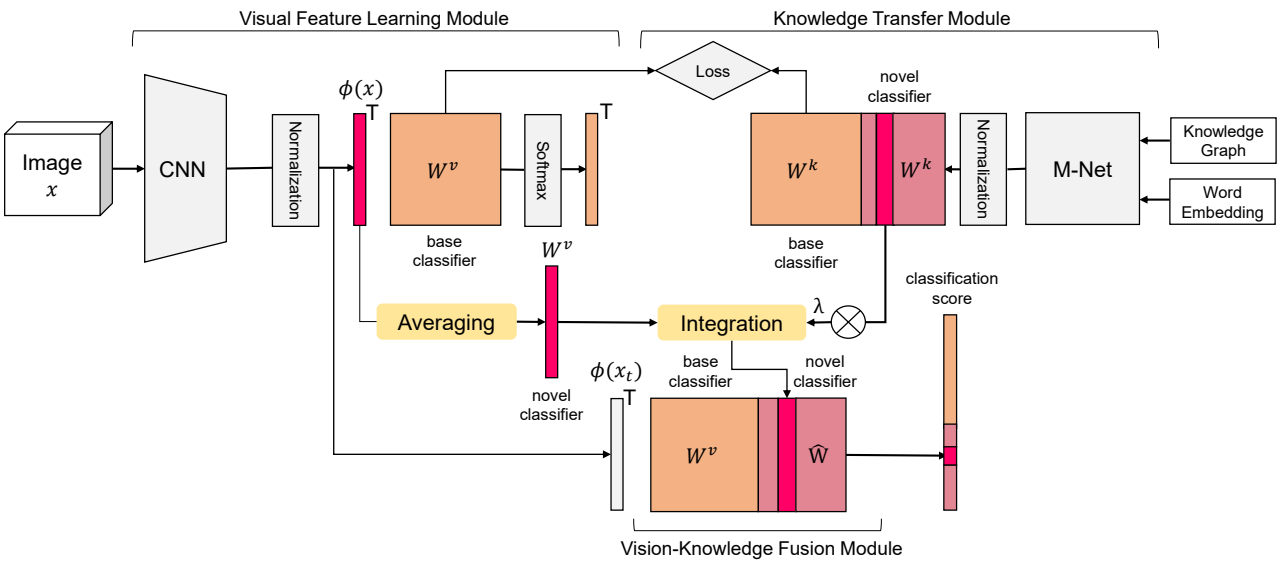

Figure 2. Illustration of the proposed Knowledge Transfer Network architecture (KTN) for few-shot image recognition.

so that models can learn new tasks quickly. These "meta knowledge" include well network initialization [3], distance metric [39] or optimizing strategy [24], etc. However, their results are still unsatisfactory because the prior knowledge has been mostly unexplored.

Actually, human vision can well recognize images of novel categories just after browsing few images of these categories. The reason may be that human vision can explore not only explicit visual information about novel objects, but also some external discriminative visual information from their prior knowledge. When they see the novel object at the next time, the visual information and inferred discriminative information will be jointly explored to make their final judgment as shown in Figure 1. Motivated by this, we propose a new few-shot learning model with deep networks by effectively exploring the explicit visual information and the implicit prior knowledge simultaneously for learning visual classifiers of novel categories.

Towards this end, we propose a novel few-shot learning method termed as Knowledge Transfer Network (KTN) by jointly incorporating visual feature learning, knowledge inferring and classifier learning into one unified framework for their optimal compatibility, as illustrated in Figure 2. It enables to adaptively leverage the explicit visual information and the implicit prior knowledge. Specifically, a visual feature extractor based on Convolutional Neural Network (CNN) [11] is trained by optimizing cosine similarity with the training data of base categories, which is used to extract the representation of examples and generate vision-based classifiers of novel categories. To well leverage the prior knowledge, a semantic-visual mapping network (M-Net) is developed to conduct knowledge inference and the semantic relationship of all categories is explicitly explored by employing the graph convolutional network [9] and knowledge graph. This mapping can serve as the knowledge-based classifiers generator of novel categories. Finally, an adaptive fusion scheme is proposed to infer the final classifiers by integrating the above two classifiers. To evaluate the effectiveness of the proposed method, extensive experiments are conducted on two widely-used Mini-ImageNet [33] and ImageNet Few-Shot [6] benchmarks. The results demonstrate the encouraging performance of proposed method compared with the state-of-the-art approaches, especially on 1 shot and 2 shot tasks.

The key contributions of this paper are summarized as follows. (1) We propose a novel Knowledge Transfer Network architecture (KTN) by jointly incorporating visual feature learning, knowledge inferring and classifier learning into one unified framework for few-shot image recognition. (2) To fully explore the prior knowledge, a semantic-visual mapping network is developed to conduct knowledge inference for novel categories from base categories. (3) An adaptive fusion scheme is proposed to infer the classifiers by integrating the visual information and knowledge information.

2. Related work

This section will briefly discuss the recent related methods to our work.

Metric-learning based Methods. Metric-learning methods try to learn an appropriate feature embedding space in which images of the same category are similar while images of different categories are dissimilar [20, 30, 14]. The results can be obtained by the nearest neighbor search.

For few-shot learning, A Siamese neural network is proposed to compute the similarity score of a pair of input images in [10]. In [33], a matching network is proposed by introducing the attention mechanism and memory unit to compare the testing and support examples. The Prototypical Network [29] takes the mean of the embedding of images in novel categories as the class prototype, and predicts the results by finding the nearest neighbor. Ren *et al*. improved the prototypical network by introducing three clustering algorithms based on semi-supervised learning [26]. Sung *et al*. considered that the current fixed metric in few-shot learning is inappropriate and proposed a Relation Network to learn a transferable deep metric [39].

Meta-learning based Methods. Meta-learning methods conduct "learning to learn" on training data to learn "meta knowledge" that can guide the rapid learning of current new tasks [2, 28, 32, 31]. For few-shot learning, a regression network is proposed to learn a generic and category agnostic transformation by regressing the few-shot classifiers and corresponding many-shot classifiers on the known categories in [37]. By observing that the procedure of gradient descent is very similar with the update procedure of LSTM [8], a LSTM meta learner is designed to learn the procedure of gradient descent in [24]. Different from them, good initial network weights are learned to be easily fine-tuned for new tasks in [3]. Mishra *et al*. proposed to aggregate information from past experiences by using temporal convolutional and soft attention [18].

Parameter generation based Methods. Parameter generation based methods can adaptively predict the classifier weights of novel categories from the feature embedding of novel examples [23, 5, 22]. In [5], an attention-based mechanism is proposed to boost the generated classifiers. Qi *et al*. applied the fine-tuning step after generating the classifiers to move the embedding space of each novel categories conforming to unimodal distribution [22]. However, all these methods predict the classifier weights only from the visual information of novel examples. The rich prior knowledge contained in the semantic embedding of category labels has not been explored. The proposed method can effectively adopt the prior knowledge to obtain some external information and incorporate it to generate more discriminative classifiers of novel categories for few-shot learning.

Zero-shot learning. Zero-shot learning (ZSL) and few-shot learning are related problems. ZSL [12, 13] aims to recognize an object instance from a new category never seen before. The category characteristic of unseen classes are learned from auxiliary prior knowledge. The commonly used prior knowledge in ZSL include human annotated attribute features of images [12, 1], the text descriptions of the image categories [25] and word embedding of the category labels [4, 19]. For more current zero-shot approaches please refer to [38]. Few-shot learning learns novel categories through only one or few examples, which lead to poor recognition accuracy. Thus it is reasonable to introduce some prior knowledge that was used in ZSL to boost the few-shot image recognition.

3. Preliminary

3.1. Problem Definition

Given a dataset that used for training a few-shot image classification model, it contains three parts: training set D_{train}, support set $D_{support}$ and testing set D_{test}. The training set has a separate category space and each category has a large amount of labeled image examples. These categories in D_{train} are defined as base categories C_{base}. Conversely, the support set $D_{support}$ and the testing set D_{test} have the same category space that are disjoint with the training set D_{train}. The categories in $D_{support}$ and D_{test} are defined as novel categories C_{novel}. If the support set contains M novel categories and each novel category has K image examples, this few-shot learning problem is defined as M-way K-shot learning. The goal of few-shot learning is to learn an image classification model by using the training set and the support set that can accurately classify images in the testing set from novel categories when K is small.

3.2. Graph Convolutional Network

We introduce Graph Convolutional Network (GCN) to learn the semantic-visual mapping by exploring the category correlation [9]. Given a graph $G = (V, E)$ and an dataset $X = \{x_j\}_{j=1}^{N}$ of N entities, each node in G is associated with a feature description, i.e., x_j is the feature vector of the identity j. The edge between two nodes in this graph G denotes their correlation. Here the number of nodes in G is N and the dimension of the feature representation is D. Thus, we have a feature matrix $X \in R^{N\times D}$ and an adjacency matrix $A \in R^{N\times N}$.

The feature matrix X and the adjacency matrix A are then integrated into a two-layer GCN simultaneously, which gives rise to the following representation of each node by combining both node content and edge correlation in the graph:

$$F = \hat{A}ReLU(\hat{A}XU_0)U_1 \tag{1}$$

where, $\hat{A}$ is the normalized A, U_0 is the weight matrix of the first layer and U_1 is the weight matrix of second layer. U_0 maps the representation of nodes to corresponding hidden states while U_1 maps the hidden state to corresponding outputs. It is noted that this graph convolution network can be extended to multiple layers to output deeper representation of graph nodes.

4. The Proposed KTN Model

In this section, we will elaborate the proposed KTN model for few-shot image recognition. The proposed ar-

chitecture contains the visual feature learning module, the knowledge transfer module and the vision-knowledge fusion module, as shown in Figure 2.

For the visual feature learning module, a CNN is trained by optimizing the cosine similarity on the whole training data D_{train}. The cosine similarity metric enables to reduce the gap between the metrics used in the training and testing phases, which lead to that the testing examples from novel categories are closer to the corresponding support examples in embedding space. And the vision-based classifiers W^v for novel categories are obtained by averaging the normalized feature embedding of novel images $\phi(x)$. For the knowledge transfer module, the input of the semantic-visual mapping network (M-Net) is the word embedding of all categories labels and a symmetric adjacency matrix encoded by the categories correlation in a knowledge graph. The M-Net is learned by maximizing the consistency between the vision-based classifiers of base categories W^v in above trained CNN and the knowledge-based classifiers of base categories W^k generated by the M-Net. Then the knowledge-based classifiers W^k for novel categories is inferred with the learned M-Net. For the vision-knowledge fusion module, an adaptive scheme is designed to learn the final classifiers $\widehat{W}$ by integrating both vision- and knowledge- based classifiers.

4.1. Visual Feature Learning Module

Given the training dataset D_{train} for base categories C_{base}, one image classification model is trained based on CNN by using all the training data in D_{train}. The traditional CNN model uses the inner-product without normalization as the metric, which leads to a gap between the metrics used in the training and testing phase [34]. To address this problem, we introduce the cosine-similarity metric to calculate the classification score in the CNN model. The cosine-similarity metric can be seen as the normalized version of inner-product, which can well improve the performance [5, 22]. With the extracted feature representation of training data $\phi(x)$ and the ℓ_2-normalized classifier W^v, the classification score s_y for a base category y can be obtained as follows:

$$s_y = \kappa(\frac{\phi(x)}{||\phi(x)||_2})^T W_y^v \tag{2}$$

where κ is the scalar parameter used to control the range of s_y, which can keep the convergence during the model training.

To have the feature representation $\phi(x)$ similar to W^v containing both positive and negative values, the ReLu nonlinearity after the last hidden layer is removed. By using the softmax cross-entropy loss function, the CNN model is trained by the following objective:

$$\min \sum_{(x,y)\in D_{train}} [-s_y + log \sum_{y'\in C_{base}} e^{s_{y'}}] \tag{3}$$

where $s_{y'}$ is the classification score for category $y' \in C_{base}$.

Once the CNN model with cosine similarity optimization is trained, the feature representations of images in $D_{support}$ can be extracted. Given the support dataset $D_{support}$ for novel categories C_{novel}, the number of images in each novel category is K. By employing the feature extractor of the CNN model with cosine similarity, the vision-based classifiers of novel categories can be inferred by averaging the normalized feature representation of the corresponding examples in $D_{support}$ [22]. That is, the vision-based classifier of a novel category y is obtained as follows:

$$W_y^v = \sum_{i=1}^{K} \phi(x_i) / \| \sum_{i=1}^{K} \phi(x_i) \|_2 \tag{4}$$

where $\{x_i\}$ are K-shot examples from the category y.

4.2. Knowledge Transfer Module

For the few-shot learning problem, it could be insufficient to infer the classifiers of novel categories by exploring only the visual information. This is because the number of examples in the support set is small, which makes it hard to directly model novel categories in embedding space. The fewer the support set samples for each novel category, the more challenging it is to recognize the category. To address this challenge, it is necessary to explore the external knowledge to supplement the vision information. Therefore, we propose to introduce the knowledge graph to train the knowledge-based classifiers augmented from vision-based counterparts as inspired by [35]. For this purpose, we propose a semantic-visual mapping network for knowledge transfer between base categories and novel categories.

The input of semantic-visual mapping network consists of a knowledge graph with the correlations between categories as its edges, and the word embeddings of category labels t as the content of its nodes. Specifically, we choose a sub-graph of WordNet [17] as this knowledge graph, which contains all categories in the 21K ImageNet data [4]. Each node in this knowledge graph represents a semantic category, and two nodes are linked if they are correlated in WordNet. The category correlation is encoded by a symmetric adjacency matrix. Then, through multiple layers of graph convolutions of the knowledge graph, the semantic-visual mapping network outputs the weight of the resultant knowledge-based classifier W_y^k for each node of novel category.

To learn a good mapping network, we aim to maximize the consistency between the vision-based classifiers of base categories and the classifiers of base categories generated by the mapping network. To better couple with the visual feature learning module with the cosine similarity optimization, the cosine similarity is introduced to measure the consistency. That is, the consistent score s_y^k for a base category

Table 1. The accuracy (%) of the proposed method using different information on 5-way few-shot learning. "Vis." and "Kno." denotes methods based on the vision-based classifier and knowledge-based classifier, respectively. "V+K" refers to method based on the integrated vision-knowledge classifier.

Model	0-shot	1-shot	5-shot
Vis.	-	54.17 ± 0.77	72.24 ± 0.57
Kno.	59.97 ± 0.74	-	-
V+K	-	$\mathbf{64.42 \pm 0.72}$	$\mathbf{74.16} \pm 0.56$

y is computed as follows:

$$s_y^k = \kappa(W_y^k)^T W_y^v \tag{5}$$

where W_y^k is the ℓ_2-normalized output of the mapping network for each node of base category and W_y^v is the ℓ_2-normalized classifier of a base category y learned by the CNN model.

By using the softmax cross-entropy loss function, the mapping network is trained by the following objective:

$$\min \sum_{(t,y)\in C_{base}} [-s_y^k + log \sum_{y'\in C_{base}} e^{s_{y'}^k}] \tag{6}$$

where $s_{y'}^k$ is the similarity score for category $y' \in C_{base}$.

4.3. Vision-Knowledge Fusion Module

The classifiers learned by exploring only the visual information is unsatisfactory due to the distribution difference of classifiers of base categories and novel categories. Especially when the number of each novel category examples in support set is small, the difference will become significant. Consequently, it is necessary to explore additional information to supplement the visual information. For this goal, prior knowledge is explored by using the proposed semantic-visual mapping network to output the knowledge-based classifiers W^k from semantic knowledge of novel categories.

Intuitively, the knowledge-based classifier and the vision-based classifier are complementary to each other. Therefore, we propose a fusion module to integrate them to obtain the final classifier $\widehat{W}$. Given a test image x_t, the few-shot image prediction is conducted as follows:

$$\begin{aligned} y^* &= \arg\max(\langle [\phi(x_t), \phi(x_t)], [W_y^v, \lambda W_y^k] \rangle) \\ &= \arg\max(\langle \phi(x_t), (W_y^v + \lambda W_y^k) \rangle) \\ &= \arg\max(\langle \phi(x_t), \widehat{W}_y \rangle) \end{aligned} \tag{7}$$

where "[ ]" is the concatenation operation, and λ is a positive balancing coefficient between two classifiers. λ is empirically set to $\frac{1}{K}$ in experiments.

Table 2. The accuracy (%) of exploring the category correlation on 5-way learning. "K_F" and "K_G" denote methods based on the knowledge-based classifiers inferred by FCN and GCN, respectively. "$V + K_F$" and "$V + K_G$" are the ones based on the corresponding integrated vision-knowledge classifiers, respectively.

Model	0-shot	1-shot	5-shot
K_F	55.03 ± 0.77	-	-
K_G	$\mathbf{59.97 \pm 0.74}$	-	-
$V + K_F$	-	62.14 ± 0.75	73.66 ± 0.56
$V + K_G$	-	$\mathbf{64.42 \pm 0.72}$	$\mathbf{74.16 \pm 0.56}$

5. Experiments

To evaluate the effectiveness of the proposed method, extensive experiments are conducted for few-shot image recognition.

5.1. Dataset

In this work, we conduct extensive experiments on two publicly available and widely used datasets: the Mini-ImageNet dataset [33] and the ImageNet Few-Shot dataset [6].

Mini-ImageNet. The Mini-ImageNet dataset is a subset of the ImageNet dataset, which contains 100 different categories with 600 images per category. The size of each image is 84×84. Following [24], the training set contains 64 categories, the validation set contains 16 categories and the testing set contains 20 categories.

ImageNet Few-Shot Dataset. The ImageNet Few-Shot dataset contains all 1000 categories in the ImageNet1K challenge. They are divided into 389 categories and 611 categories for base categories and novel categories, respectively. Images from 193 categories of the base categories and 300 categories of the novel categories are used for cross validation. Images of the remaining 196 base categories and 311 novel categories are used for testing (please refer to [6] for more details).

5.2. Experimental Setting

For the Mini-ImageNet dataset, following previous few-shot learning approaches [33, 24, 3, 29, 39, 5], we utilize a four layer CNN (ConvNet) in which each convolutional block has 64 filters (64F) and the size of all filters is 3×3. For fair comparison, we also employ another four layer CNN in which the first two convolutional layers have 64 filters and the latter two convolutional layers have 128 filters (128F) and a ResNet that used in previous works [5, 18]. For the ImageNet Few-Shot dataset, the ResNet-10 and ResNet-50 are used by following [36]. For CovNet, the first three convolutional blocks are set with batch normalization, ReLu non-linearity and 2×2 max-pooling, the last convolutional block only with batch normalization and 2×2 max-pooling, respectively. For the knowledge trans-

Table 3. The average classification accuracies (%) with 95% confidence intervals on the Mini-ImageNet dataset.

Model	Feature Extractor	5-way 1-shot	5-way 5-shot
Matching Networks [33]	ConvNet(64F)	43.56 ± 0.84	55.31 ± 0.73
Meta-learner LSTM [24]	ConvNet(32F)	43.44 ± 0.77	60.60 ± 0.71
MAML [3]	ConvNet(64F)	48.70 ± 1.84	63.11 ± 0.92
Prototypical-Nets [29]	ConvNet(64F)	49.42 ± 0.78	68.20 ± 0.66
Relation Net [39]	ConvNet(64F)	50.44 ± 0.82	65.32 ± 0.70
SNAIL [18]	ResNet	55.71 ± 0.99	68.88 ± 0.92
DFVL [5]	ConvNet(64F)	56.20 ± 0.86	72.81 ± 0.62
DFVL [5]	ConvNet(128F)	55.95 ± 0.71	73.00 ± 0.64
DFVL [5]	ResNet	55.45 ± 0.89	70.13 ± 0.68
Ours(Vis.)	ConvNet(64F)	54.61 ± 0.80	71.21 ± 0.66
Ours(V+K)	ConvNet(64F)	64.06 ± 0.72	73.27 ± 0.54
Ours(Vis.)	ConvNet(128F)	54.17 ± 0.77	72.24 ± 0.57
Ours(V+K)	ConvNet(128F)	$\mathbf{64.42 \pm 0.72}$	$\mathbf{74.16 \pm 0.56}$
Ours(Vis.)	ResNet	54.34 ± 0.77	69.02 ± 0.65
Ours(V+K)	ResNet	61.42 ± 0.72	70.19 ± 0.62

fer module, we use all the categories in the ImageNet 21K dataset and their correlations in WordNet to construct the knowledge graph.

For the Mini-ImageNet datset, the CNN-based feature extractor is trained for 60 epochs. The parameters are learned by using stochastic gradient descent with a mini-batch size of 256. For the ImageNet Few-Shot dataset, the CNN-based feature extractor is trained for 100 epochs. The parameters are learned by using stochastic gradient descent with a mini-batch size of 400 for ResNet-10 and 160 for ResNet-50 respectively. The initial learning rate is set to 0.1 for ResNet-10 and 0.025 for ResNet-50 respectively. The weight decay is set to 0.0005 and the momentum is set to 0.9. The semantic-visual mapping network (M-Net) consists three layers. The numbers of nodes in the hidden layer and the output layer are the same to the size of W^v. Leaky ReLu is used with a negative slope of 0.2. Each layer of the mapping network is followed with a Dropout operation with the rate of 0.5. The mapping network is trained with 20 and 250 epochs for Mini-ImageNet dataset and ImageNet Few-Shot dataset respectively. The Glove embedding model [21] trained on the Wikipedia dataset is introducted for word embedding of each category and the dimension of the word embedding is 300. The learning rate is set to 0.001 and the weight decay is set to 0.0005. The Adam optimizer is used for training. All the scale parameter κ in experiments are set to 10.

The classification accuracy is used to evaluate the performance of the few-shot learning methods. For the Mini-ImageNet dataset, following previous few-shot learning approaches, experiments for 5-way 1-shot and 5-way 5-shot image classification tasks are conducted. There are 15 testing images for each novel category. For the ImageNet Few-Shot dataset, the evaluation protocol is the same to the one in [36]. The top-5 classification accuracy is computed. Following [36], three evaluation criteria are used: "Novel" (the testing images from C_{novel} and the label from C_{novel}), "All" (the testing images from C_{base} and C_{novel} in equal proportion as the label from C_{base} and C_{novel}), and "All with prior" (the testing images from C_{base} and C_{novel} as the label from C_{base} and C_{novel} with a novel class prior).

5.3. Ablation Study

In this section, we conduct an ablation study on the Mini-ImageNet dataset.

First, experiments are carried out to verify the effectiveness of fusing different information. Two variants of the proposed model by only exploring the visual information and knowledge information respectively are compared. The compared results are shown in Table 1. The ConvNet (128F) is used as the feature extractor. For convenience of comparison, the performance of the knowledge-based classifier for 0-shot learning is reported. From the results, it can be observed that the proposed model achieves best results by jointly considering the visual and knowledge information. Compared to the model only using the vision-based classifier, the proposed method has the significant improvement about 10% for the 5-way 1-shot classification task and about 2% for the 5-way 5-shot classification task. It well indicates the effectiveness and necessity of exploring external knowledge. Besides, the proposed method outperforms the model with only prior knowledge, which shows the importance of the visual information. Finally, the improvement decreases when the number of images in the support set increases. It may be because that the distribution difference becomes less with more images in the support sets. In short,

Table 4. The results of top-5 average classification accuracies on the ImageNet Few-Shot dataset. "w/A" and "w/G*" mean using hallucinated additional examples for novel categories.

	Novel					All					All with prior				
Method	n=1	2	5	10	20	n=1	2	5	10	20	n=1	2	5	10	20
ResNet-10															
PN [29]	39.3	54.4	66.3	71.2	73.9	49.5	61.0	69.7	72.9	74.6	53.6	61.4	68.8	72.0	73.8
MN [33]	43.6	54.0	66.0	72.5	76.9	54.4	61.0	69.7	73.7	76.5	54.5	60.7	68.2	72.6	75.6
LogReg [36]	38.4	51.1	64.8	71.6	76.6	40.8	49.9	64.2	71.9	76.9	52.9	60.4	68.6	72.9	76.3
LogReg w/A [36]	40.7	50.8	62.0	69.3	76.5	52.2	59.4	67.6	72.8	76.9	53.2	59.1	66.8	71.7	76.3
PMN* [36]	43.3	55.7	68.4	74.0	77.0	55.8	63.1	71.1	75.0	77.1	54.7	62.0	70.2	73.9	75.9
PMN w/G* [36]	45.8	57.8	69.0	74.3	77.4	57.6	64.7	71.9	75.2	77.5	56.4	63.3	70.6	74.0	76.2
DFVL Avg. [5]	45.23	59.60	68.68	74.36	77.69	57.65	64.69	72.35	76.18	78.46	56.43	63.41	70.95	74.75	77.00
DFVL Att. [5]	46.02	57.51	69.16	74.83	**78.11**	58.16	65.21	72.72	**76.50**	**78.74**	56.76	63.80	**72.72**	**75.02**	**77.25**
Ours (Vis.)	45.44	56.71	68.91	74.50	77.71	55.90	63.34	71.89	76.08	78.31	56.37	63.20	70.86	74.64	76.77
Ours (V+K)	**54.74**	**61.69**	**70.36**	**74.98**	77.86	**62.08**	**66.79**	**73.08**	76.44	78.44	**61.71**	**66.07**	71.78	74.92	76.91
ResNet-50															
MN [33]	53.5	63.5	72.7	77.4	81.2	64.9	71.0	77.0	80.2	82.7	63.8	69.9	75.9	79.3	81.9
PN [29]	49.6	64.0	74.4	78.1	80.0	61.4	71.4	78.0	80.0	81.1	62.9	70.5	77.1	79.5	80.8
PN w/G* [29]	53.9	65.2	75.7	80.2	82.8	65.2	72.0	78.9	81.7	83.1	63.9	70.5	77.5	80.6	82.4
PMN* [36]	53.3	65.2	75.9	80.1	82.6	64.8	72.1	78.8	81.7	83.3	63.4	70.8	77.9	80.9	82.7
PMN w/G* [36]	54.7	66.8	**77.4**	**81.4**	**83.8**	65.7	73.5	**80.2**	**82.8**	**84.5**	64.4	71.8	**78.7**	**81.5**	**83.3**
Ours (Vis.)	53.6	65.2	75.5	79.8	82.3	64.8	71.8	78.9	81.9	83.6	63.6	70.7	77.7	80.7	82.4
Ours (V+K)	**61.9**	**68.7**	76.4	80.1	82.4	**69.7**	**74.1**	79.4	82.0	83.7	**68.6**	**73.0**	78.3	80.9	82.5

it is necessary to explore the external knowledge as supplement of the visual information.

Next, experiments are conducted to show the effectiveness of the explicit category correlation for knowledge-based classifier learning. For fair comparison, we employ a fully connection network (FCN) as the semantic-visual mapping network that does not explore the category correlation. It can serve as a naïve baseline method that exploring both visual and knowledge information. The compared results of the mapping network based on GCN and FCN are represented in Table 2. It can be seen that the semantic-visual mapping network based on GCN is superior to the one based on FCN, which shows the importance of exploring the explicit category correlations in knowledge graph. More category-wise detailed information can be uncovered by introducing GCN for the semantic-visual mapping.

5.4. Experimental Results and Analysis

This section discusses the results of the proposed method on the two datasets.

First, experiments are conducted on the Mini-ImageNet dataset. The proposed KTN is compared with several state-of-the-art few-shot learning approaches, including Matching Networks [33], Meta-learner LSTM [24], MAML [3], SNAIL [18], Prototypical Nets [29], Relation Network [39] and Dynamic Few-shot Visual Learning (DFVL) [5]. The compared results in terms of the average classification accuracy with 95% confidence intervals are demonstrated in Table 3. Experiments are independently repeated 600 times and the testing data is randomly sampled, the average results are reported. It can be seen that the proposed KTN achieves the best performance for 5-way 1-shot and 5-way 5-shot image recognition tasks. This result indicates that it is effective to incorporate the visual feature learning, knowledge inferring and classifier learning in a unified framework, especially when the number of examples in the support set is very small. It can well verify the motivation of the proposed method. By comparing the results of DFVL and KTN with different feature extractors, the proposed model with 64 filters gains better results than DFVL with 128 filters on both 5-way 1-shot task and 5-way 5-shot tasks, which well shows the effectiveness of the proposed method.

For the ImageNet Few-Shot dataset, we conduct experiments to compare the KTN with several previous methods, including Matching Networks [33], Prototypical Nets [29], Logistic regression [36], Prototype Matching Nets (PMN) [36] and DFVL [5]. Experiments are independently repeated 100 times with randomly sampled testing data, and the average results are reported. The results with 95% confidence intervals are shown in Table 4. All the results of the compared methods are from [5] and [36]. We can observe that the KTN achieves the best or competitive perfor-

Table 5. The results of top-5 average accuracies in terms of AFNE on the ImageNet Few-Shot dataset.

Model	AFNE				
	N=1	2	5	10	20
DFVL [5]	40.68	51.61	63.75	**70.09**	**74.00**
Ours(Vis.)	34.26	46.46	61.55	69.19	73.63
Ours(V+K)	**45.35**	**53.22**	**63.98**	**70.09**	73.93

mance. Specially, the proposed method gains the significant improvement on the 1-shot and 2-shot image recognition tasks. It well demonstrate the effectiveness of the proposed method by leveraging the external knowledge information. When the number of samples in the support set become larger, the advantage of the proposed KTN is inconspicuous because more examples in the support set may provide sufficient information for classification. In brief, the proposed KTN achieves very encouraging results by introducing the knowledge transfer.

Finally, since this work focuses on the problem of few-shot image recognition, we should only pay attention to the performance of testing examples from C_{novel}. However, for the ImageNet Few-Shot dataset, the results of two evaluation criteria ("All" and "All with prior") are computed to evaluate the ability to not forget the base categories by sampling the testing examples from both C_{base} and C_{novel}. We assume that it can lead to that the performance of testing examples from C_{base} would blood the performance of testing examples from C_{novel}. To address this problem, we conduct experiments by using a new evaluation criteria that the testing examples are only from C_{novel} and the true labels are from both C_{base} and C_{novel}. It can evaluate the ability to not forget the base categories more reasonably. This criteria is termed as All For Novel categories Examples (AFNE) in this work. The results of top-5 average classification accuracies in terms of AFNE on the ImageNet Few-Shot dataset are shown in Table 5. From Table 4 and Table 5, we can observe the results in terms of AFNE are significantly lower than the results in terms of "All" and "All with prior", which well verifies our assumption. In the meanwhile, the results show that the proposed method achieves bigger improvement over the one without the knowledge transfer, which can better demonstrate the effectiveness of exploring the external knowledge information.

5.5. Visualizing the Fused Weight

To well show the importance of the introduced knowledge information, we perform t-SNE visualization to present the classifiers learned by the proposed model and the one without considering the knowledge information. The results of all 20 novel categories on the Mini-ImageNet testing set for the 1-shot task and the 5-shot task are illustrated in Figure 3. It can be easily observed that the clustering results by incorporating the visual information and the knowledge information are more compact than the ones by only considering the visual information, which can well illustrate the discriminative ability of the proposed model.

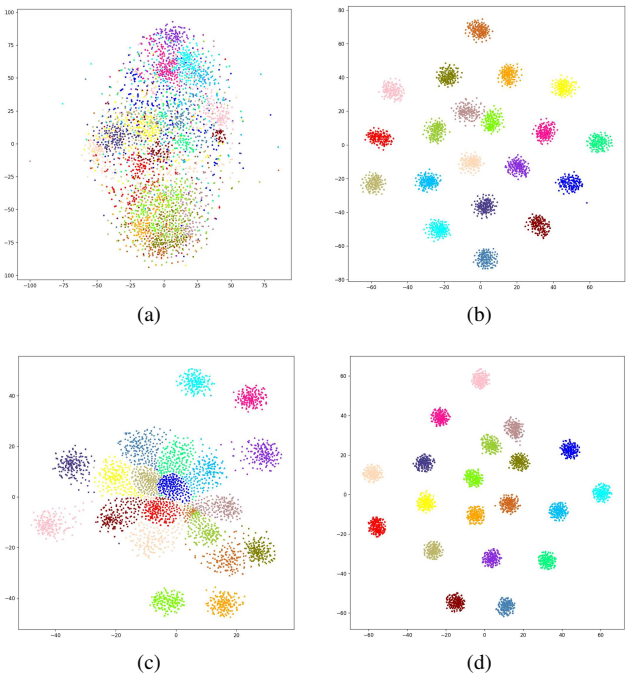

Figure 3. T-SNE visualization results for all novel categories in the Mini-imagenet set on the 1-shot and 5-shot tasks. Each scatter plot contains 20 colored classifier parameter clusters and each color represents a novel category. (a): 1-shot vision-based classifier. (b): 1-shot vision-knowledge classifier. (c): 5-shot vision-based classifier. (d): 5-shot vision-knowledge classifier.

6. Conclusion

In this paper, we propose a novel Knowledge Transfer Network architecture (KTN) by jointly incorporating visual feature learning, knowledge inferring and classifier learning into one unified framework for few-shot image recognition. To well explore the external knowledge information, a semantic-visual mapping network based on GCN is developed for knowledge transfer. The visual information and the knowledge information are fused to learn the final classifier. Experimental results on two publicly available datasets show the encouraging performance of the proposed method.

7. Acknowledgments

This work was partially supported by the National Key Research and Development Program of China under Grant 2017YFC0820601, the National Natural Science Foundation of China (Grant No. 61772275, 61732007, 61720106004 and 61876181) and the Natural Science Foundation of Jiangsu Province (BK20170033).

References

[1] Zeynep Akata, Florent Perronnin, Zaïd Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *TPAMI*, 2016.

[2] Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando De Freitas. Learning to learn by gradient descent by gradient descent. In *NIPS*, 2016.

[3] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017.

[4] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013.

[5] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *CVPR*, 2018.

[6] Bharath Hariharan and Ross Girshick. Low-shot visual recognition by shrinking and hallucinating features. In *ICCV*, 2017.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[8] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 1997.

[9] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017.

[10] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML Workshop*, 2015.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[12] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009.

[13] Yan Li, Junge Zhang, Jianguo Zhang, and Kaiqi Huang. Discriminative learning of latent features for zero-shot recognition. In *CVPR*, 2018.

[14] Zechao Li and Jinhui Tang. Weakly supervised deep metric learning for community-contributed image retrieval. *IEEE Transactions on Multimedia*, 2015.

[15] Zechao Li, Jinhui Tang, and Tao Mei. Deep collaborative embedding for social image understanding. *IEEE TPAMI*, 2018.

[16] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[17] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 1995.

[18] Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A simple neural attentive meta-learner. In *ICLR*, 2018.

[19] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. In *ICLR*, 2014.

[20] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *CVPR*, 2016.

[21] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.

[22] Hang Qi, Matthew Brown, and David G Lowe. Low-shot learning with imprinted weights. In *CVPR*, 2018.

[23] Siyuan Qiao, Chenxi Liu, Wei Shen, and Alan L. Yuille. Few-shot image recognition by predicting parameters from activations. In *CVPR*, 2018.

[24] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *ICLR*, 2016.

[25] Scott Reed, Zeynep Akata, Bernt Schiele, and Honglak Lee. Learning deep representations of fine-grained visual descriptions. In *CVPR*, 2016.

[26] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. Meta-learning for semi-supervised few-shot classification. In *ICLR*, 2018.

[27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[28] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *ICML*, 2016.

[29] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NIPS*, 2017.

[30] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *NIPS*, 2016.

[31] Sebastian Thrun. Lifelong learning algorithms. In *Learning to learn*, pages 181–209. 1998.

[32] Sebastian Thrun and Lorien Pratt. Learning to learn: Introduction and overview. In *Learning to learn*, pages 3–17. 1998.

[33] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *NIPS*, 2016.

[34] Feng Wang, Xiang Xiang, Jian Cheng, and Alan Loddon Yuille. Normface: l 2 hypersphere embedding for face verification. In *ACM MM*, 2017.

[35] Xiaolong Wang, Yufei Ye, and Abhinav Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. In *CVPR*, 2018.

[36] Yu-Xiong Wang, Ross Girshick, Martial Herbert, and Bharath Hariharan. Low-shot learning from imaginary data. In *CVPR*, 2018.

[37] Yu-Xiong Wang and Martial Hebert. Learning to learn: Model regression networks for easy small sample learning. In *ECCV*, 2016.

[38] Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. Zero-Shot Learning – A Comprehensive Evaluation of the Good, the Bad and the Ugly. *TPAMI*, 2018.

[39] Flood Sung Yongxin Yang, Li Zhang, Tao Xiang, Philip H-S Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *CVPR*, 2018.