

# Knowledge Graph Transfer Network for Few-Shot Recognition

Riquan Chen,<sup>1</sup> Tianshui Chen,<sup>1,2</sup> Xiaolu Hui,<sup>1</sup> Hefeng Wu,<sup>1\*</sup> Guanbin Li,<sup>1</sup> Liang Lin<sup>1,2</sup>

<sup>1</sup>Sun Yat-sen University, <sup>2</sup>DarkMatter AI Research  
 sysucrq@gmail.com, tianshuichen@gmail.com, huixlu@mail2.sysu.edu.cn, wuhefeng@gmail.com  
 liguanbin@mail.sysu.edu.cn, linlng@mail.sysu.edu.cn

## Abstract

Few-shot learning aims to learn novel categories from very few samples given some base categories with sufficient training samples. The main challenge of this task is the novel categories are prone to be dominated by color, texture, shape of the object or background context (namely *specificity*), which are distinct for the given few training samples but not common for the corresponding categories (see Figure 1). Fortunately, we find that transferring information of the correlated based categories can help learn the novel concepts and thus avoid the novel concept being dominated by the *specificity*. Besides, incorporating semantic correlations among different categories can effectively regularize this information transfer. In this work, we represent the semantic correlations in the form of structured knowledge graph and integrate this graph into deep neural networks to promote few-shot learning by a novel Knowledge Graph Transfer Network (KGTN). Specifically, by initializing each node with the classifier weight of the corresponding category, a propagation mechanism is learned to adaptively propagate node message through the graph to explore node interaction and transfer classifier information of the base categories to those of the novel ones. Extensive experiments on the ImageNet dataset show significant performance improvement compared with current leading competitors. Furthermore, we construct an ImageNet-6K dataset that covers larger scale categories, i.e., 6,000 categories, and experiments on this dataset further demonstrate the effectiveness of our proposed model.

## Introduction

Recently, deep convolutional neural networks (ConvNet) (Krizhevsky, Sutskever, and Hinton 2012; He et al. 2016) have obtained remarkable success in various visual recognition tasks. To fully train a deep ConvNet recognition system, it is requested that each category has thousands of training samples. If the system needs to recognize some novel categories, we need to collect large amounts of training samples for these categories to avoid being overfitting. In contrast,

\*Riquan Chen and Tianshui Chen contribute equally and share first-authorship. Corresponding author is Hefeng Wu. This work is partly supported by National Natural Science Foundation of China (No. U1611461 and 61876045) and SenseTime Research Fund. Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

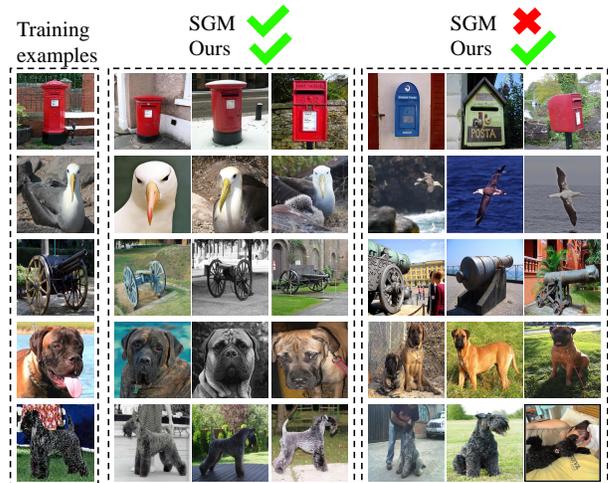


Figure 1: Visualization results of SGM and our proposed models. The first column shows the only training samples of different categories. The next three columns show the examples that are correctly classified by both two models. The last three columns show the samples that are misclassified by the SGM model but correctly classified by our model. It can be observed that the SGM model can well classify samples with high similarity to the training image but fails in the samples with a large difference in appearance. In contrast, our proposed model shows superior performance on more appearance patterns.

human can effortlessly learn to recognize novel categories with few samples by exploiting prior knowledge accumulated from daily life. Mimicking this ability to learn novel categories from very few samples (also known as few-shot learning) is a crucial yet practical task in the computer vision community.

Formally, we consider a general and practical scenario for few-shot learning, in which there exist a set of base categories with sufficient training samples and a set of novel categories with very limited samples (e.g., no more than ten samples in this work). Unlike previous works that merely focus on the novel categories, we aim to develop a recog-

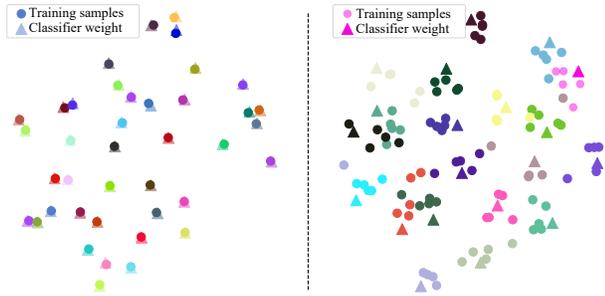


Figure 2: The t-SNE result of the normalized training samples and classifier weight. Scatter plot from the left to the right shows 1-shot and 5-shot setting, respectively.

dition system that learns to recognize the novel categories based on few training samples and simultaneously maintain the performance on the base categories.

It is well-known that each image contains features that are particular and discriminant for this image but not common for its category (e.g., texture, color, shape of the object or background context). Here, we denote these features as *specificity*. In few-shot learning scenarios, each novel category has very limited samples that can hardly describe its commonality, so its learned classifier is prone to be dominated by the *specificity* and thus may deviate severely from reality. Take the *Kerry blue terrier* in the last row of Figure 1 for example, which is in the 1-shot setting on ImageNet-FS dataset, the state-of-the-art model SGM (Hariharan and Girshick 2017) can well classify the instances with a specific side-standing pose but fail in instances with other poses. To delve deep into this phenomenon, we further visualize the extracted features of training samples and the corresponding classifier weights learned by SGM (Hariharan and Girshick 2017) in Figure 2. As shown, the learned classifier is highly correlated with the extracted features. If the training samples of a category mainly contain *specificity*, the learned classifier may squint towards describing this *specificity* and inevitably miss the common feature for this category.

Fortunately, strong semantic correlations are witnessed among categories with similar visual concepts, which originates from human cognitive system. Transferring the information of the correlated base categories can provide additional information to guide learning the novel concept and thus avoid the novel concept being dominated by the *specificity*. Moreover, these category correlations can effectively regularize this information transfer. In this work, we develop a Knowledge Graph Transfer Network (KGTN) that integrates category correlations into a deep neural network to guide exploiting information of base categories to help to learn the novel concept. To this end, we represent the semantic correlations with a knowledge graph, where each node refers to the classifier weight of a category and each edge represents the semantic correlation between the two corresponding categories. Then, we introduce a graph propagation mechanism to transfer node message through the graph. In this way, it allows each category to derive information of the correlated categories to better learn the classifier under

the explicit guidance of category correlations. Notably, the graph contains both novel and base categories and the message propagation mechanism is shared across all node pairs in the graph, and thus such a mechanism can be trained using sufficient samples of the base categories and well generalizes to the novel categories.

The contributions are summarized into three folds: 1) We propose to integrate category correlations as prior knowledge to regularize and guide transferring information of classifier weights. 2) We introduce a graph update mechanism that propagates the node message through the graph to iteratively update the classifier weights. 3) We conduct experiments on the widely used ImageNet based few-shot dataset and demonstrate the superior performance of our proposed framework over existing state-of-the-art methods. To evaluate the performance on larger scale categories, we construct an ImageNet-6K dataset that covers 6,000 categories. Experiments conducted on this benchmark shows that our model still outperforms current leading methods.

## Related Works

**Few-shot learning.** Previous works (Finn, Abbeel, and Levine 2017; Li et al. 2017; Vinyals et al. 2016; Garcia and Bruna 2017; Snell, Swersky, and Zemel 2017) follow a simple  $N$ -way  $K$ -shot setting, where there are  $N$  categories with  $K$  training samples for each category, and  $N$  is no more than 20. These works often adopt learning-to-learn paradigm that distills knowledge learned from training categories to help learn novel concepts. For example, (Vinyals et al. 2016; Snell, Swersky, and Zemel 2017; Sung et al. 2018; Kim et al. 2019) learn an embedding and metric function from the base categories to well recognize samples in the novel categories. Most of these works evaluate their algorithms on some small-scale datasets, e.g., miniImageNet with 64 base categories, 16 validation categories, 20 novel categories. Recently, some works (Hariharan and Girshick 2017; Chen et al. 2019c; Hui, Chen, and Chen 2019) switch to a more general and practical setting, where the algorithms aim to recognize hundreds of novel concepts with very limited samples given a set of based categories with sufficient training samples. To address this few-shot learning scenario, (Hariharan and Girshick 2017) learn a transformation function to hallucinate additional samples for novel categories. (Chen et al. 2019c) further consider to enlarge the size of dataset and learn a deformation network to deform images by fusing a pair of images. (Qiao et al. 2018b) explore the relation of training feature and classifier weight and adapt a neural network to obtain classifier weights directly from the training features. To evaluate these algorithms, researchers construct a larger-scale dataset that covers 1,000 categories (Hariharan and Girshick 2017). As this is a more practical scenario, we focus on this setting in our work.

**Knowledge Embedded Visual Reasoning.** Recently, lots of works attempt to incorporate prior knowledge with deep representation learning for various visual reasoning tasks, ranging from visual recognition/detection (Wang, Ye, and Gupta 2018; Chen et al. 2018c; Lee et al. 2018; Chen et al. 2019a; Chen et al. 2018b; Jiang et al. 2018) to visual relationship

reasoning (Chen et al. 2019b; Wang et al. 2018b) and navigation/planning (Yang et al. 2019; Chen et al. 2018a). As a pioneering work, (Marino, Salakhutdinov, and Gupta 2017) build a knowledge graph to correlate object categories and learn graph representation to enhance image representation learning. (Chen et al. 2019a) incorporate a similar knowledge graph to guide feature decoupling and interaction to further facilitate multi-label image recognition. (Wang, Ye, and Gupta 2018) apply graph convolutional network (Kipf and Welling 2016) to explore semantic interaction and direct map semantic vectors to classifier parameters. To further explore high-level tasks, (Chen et al. 2019b) consider the correlations between specific object pairs and their corresponding relationships to regularize scene graph generation and thus alleviate the effect of the uneven distribution issue. (Chen et al. 2018a) build And-Or Graphs (Zhu, Mumford, and others 2007) to describe tasks, which helps regularize atomic action sequence generation to reduce the dependency on annotated samples. In few-shot learning scenario, the latest work (Li et al. 2019) construct category hierarchy by semantic cluster and regularize prediction at each hierarchy via a hierarchical output net. Different from these works, we formulate classifier weight as a prototype representation of the corresponding category, and introduce a graph propagation mechanism to transfer prototype representation of base categories to guide learning novel concepts under the explicit guidance of prior semantic correlation.

## Methodology

We first revisit the few-shot problem formulation in the conventional deep ConvNet paradigm with some formalized interpretation on why such paradigm easily fails in the few-shot scenarios, and then present our model in detail.

### Problem Formulation Revisited

A few-shot recognition system should be able to recognize novel categories with limited few samples. Recent deep ConvNet models have achieved remarkable success in image recognition, but they are also notorious for requiring a large number of training samples for each category and are not fit for the few-shot scenarios. In this work, we amend the conventional paradigm to adapt deep ConvNet models to such scenarios.

A typical deep ConvNet-based image recognition model consists of the feature extractor and classifiers, which are jointly learned for optimal performance. Given an image sample  $x$ , the recognition model predicts its label  $\hat{y}$  as  $\hat{y} = \arg \max_k p(y = k|x)$ , and

$$p(y = k|x) = \frac{\exp(f_k(\mathbf{x}))}{\sum_{i=1}^K \exp(f_i(\mathbf{x}))} \quad (1)$$

is the softmax function. Here  $K$  is the number of categories, and  $\mathbf{x}$  is the  $d$ -dimensional representation feature of  $x$  output by the feature extractor  $\phi(\cdot)$ , i.e.,  $\mathbf{x} = \phi(x)$ . The linear classifier  $f_k(\mathbf{x}) = \mathbf{w}_k^\top \mathbf{x} + b_k$  computes the confidence of the sample  $x$  belonging to category  $k$ , and is implemented with a fully-connected layer.  $\mathbf{w}_k$  denotes the classifier weight and  $b_k$  is the bias term. It can be easily inferred that  $\arg \max_k p(y = k|x)$  is equivalent to  $\arg \max_k f_k(\mathbf{x})$ .

We reformulate  $f_k(\mathbf{x})$  as follows:

$$\begin{aligned} f_k(\mathbf{x}) &= \mathbf{w}_k^\top \mathbf{x} + b_k \\ &= -\frac{1}{2} \|\mathbf{w}_k - \mathbf{x}\|_2^2 + \frac{1}{2} \|\mathbf{w}_k\|_2^2 + \frac{1}{2} \|\mathbf{x}\|_2^2 + b_k \end{aligned} \quad (2)$$

By introducing the constraints  $b_k = 0$  and  $\|\mathbf{w}_i\|_2 = \|\mathbf{w}_j\|_2$ ,  $\forall i, j$ , the classifier  $f_k(\mathbf{x})$  can be viewed as a similarity measure between  $\mathbf{x}$  and  $\mathbf{w}_k$ , and we have

$$\hat{y} = \arg \max_k f_k(\mathbf{x}) = \arg \min_k \|\mathbf{w}_k - \mathbf{x}\|_2^2 \quad (3)$$

Therefore, the weight  $\mathbf{w}_k$  can be viewed as the prototype representation of category  $k$ , and the sample  $x$  is predicted as category  $k$  if its feature  $\mathbf{x}$  has the maximum similarity (or minimum distance) with prototype  $\mathbf{w}_k$ . In this perspective, we can implement  $f_k(\mathbf{x})$  with different similarity metrics.

Relaxing the above constraints to general deep ConvNet models, the prototype representation perspective of classifier weight is reasonable to some extent. Thus, when these models are applied in the few-shot scenarios, the learned prototype will be guided to reflect the *specificity* of the few training samples (as visualized in Figure 2) and cannot capture the commonality of the corresponding category.

To tackle this problem, we amend the conventional deep ConvNet paradigm by incorporating category correlations to transfer prototype presentations among similar categories and thus substantially enhance the prototypes of novel categories. Specifically, the category correlations are modeled as a knowledge graph and integrated into the deep ConvNet to build our Knowledge Graph Transfer Network model, which will be detailed below.

### Knowledge Graph Transfer Network

The overall framework of our Knowledge Graph Transfer Network (KGTN) is illustrated in Figure 3, which consists of three modules: Feature Extraction, Knowledge Graph Transfer, and Prediction. The key design is the knowledge graph transfer module where we incorporate a graph neural network to explore the knowledge transfer of the prototypes (classifier weights) by the guidance of semantic correlations on top of the ConvNet.

**Knowledge Graph Transfer Module.** We model the classifier weight into a graph, in which nodes refer to the classifier weight and edges represent their semantic correlations. Then, we incorporate a Gated Graph Neural Network (GGNN) to update and propagate the message between nodes.

Given a dataset that covers  $K = K_{base} + K_{novel}$  categories ( $K_{base}$  and  $K_{novel}$  denote the number of base and novel categories), we use a graph  $\mathcal{G} = \{\mathbf{V}, \mathbf{A}\}$  to encode the correlations among all categories, in which nodes refer to the categories and edges represent their correlations. Specifically,  $\mathbf{V}$  is represented as  $\{v_1, v_2, \dots, v_{K_{base}}, \dots, v_K\}$  where node  $v_k$  corresponds to category  $k$ , and  $\mathbf{A}$  is an adjacency matrix, in which  $a_{ij}$  denotes the correlation weight between categories  $i$  and  $j$ .

At each iteration  $t$ , node  $v_k$  has a hidden state  $\mathbf{h}_k^t$ , and the hidden state  $\mathbf{h}_k^0$  at iteration  $t = 0$  is set as the initial classifier weight  $\mathbf{w}_k^{init}$ , formulated as

$$\mathbf{h}_k^0 = \mathbf{w}_k^{init}, \quad (4)$$

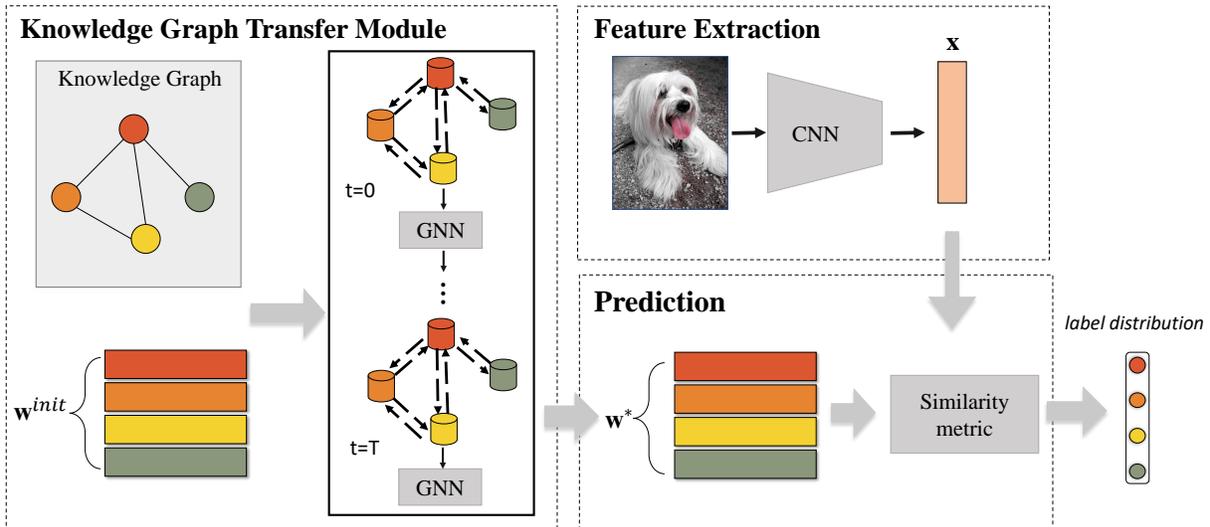


Figure 3: Illustration of the proposed Knowledge Graph Transfer Network (KGTN) model. It incorporates prior knowledge of category correlations to help explore the interactions among classifier weights of all categories to better learn the classifier weights of the novel categories.

in which  $\mathbf{w}_k^{init}$  is learnable parameters and randomly initialized before training. At each iteration  $t$ , each node  $k$  aggregates message from its correlated node such that the parameter vectors of these nodes can help refine its parameter vector, formulated as

$$\mathbf{a}_k^t = \left[ \sum_{k'=1}^K a_{kk'} \mathbf{h}_{k'}^{t-1}, \sum_{k'=1}^K a_{k'k} \mathbf{h}_{k'}^{t-1} \right]. \quad (5)$$

In this way, a high correlation between nodes  $k$  and  $k'$  encourage message propagation from  $k'$  to  $k$ , and it suppresses the propagation otherwise. Then, the framework takes this aggregated feature vector and the hidden state of the previous iteration as input to update the corresponding hidden state by a gating mechanism

$$\begin{aligned} \mathbf{z}_k^t &= \sigma(\mathbf{W}^z \mathbf{a}_k^t + \mathbf{U}^z \mathbf{h}_k^{t-1}) \\ \mathbf{r}_k^t &= \sigma(\mathbf{W}^r \mathbf{a}_k^t + \mathbf{U}^r \mathbf{h}_k^{t-1}) \\ \widetilde{\mathbf{h}}_k^t &= \tanh(\mathbf{W} \mathbf{a}_k^t + \mathbf{U}(\mathbf{r}_k^t \odot \mathbf{h}_k^{t-1})) \\ \mathbf{h}_k^t &= (1 - \mathbf{z}_k^t) \odot \mathbf{h}_k^{t-1} + \mathbf{z}_k^t \odot \widetilde{\mathbf{h}}_k^t \end{aligned} \quad (6)$$

where  $\sigma(\cdot)$  and  $\tanh(\cdot)$  are the logistic sigmoid and hyperbolic tangent functions, and  $\odot$  is the element-wise multiplication operation. In this way, the model tends to adopt the more correlated message to update parameters of the current node. The propagation is repeated by  $T$  iterations, and we can obtain the final hidden states, i.e.,  $\{\mathbf{h}_1^T, \mathbf{h}_2^T, \dots, \mathbf{h}_K^T\}$ . Finally, we utilize a simple output network to predict the classifier weight

$$\mathbf{w}_k^* = o(\mathbf{h}_k^T, \mathbf{h}_k^0). \quad (7)$$

**Prediction with different similarity metrics.** As stated in the problem formulation, the classifier  $f_k(\mathbf{x})$  can be implemented as similarity metric. Here, we consider three similarity metrics for evaluation: *inner product*, *cosine similarity*, and *Pearson correlation coefficient*.

**Inner product:**

$$f_k(\mathbf{x}) = \mathbf{x} \cdot \mathbf{w}_k^* \quad (8)$$

**Cosine similarity:**

$$f_k(\mathbf{x}) = \frac{\mathbf{x} \cdot \mathbf{w}_k^*}{\|\mathbf{x}\|_2 \cdot \|\mathbf{w}_k^*\|_2} \quad (9)$$

**Pearson correlation coefficient:**

$$f_k(\mathbf{x}) = \frac{(\mathbf{x} - \bar{\mathbf{x}}) \cdot (\mathbf{w}_k^* - \bar{\mathbf{w}}_k^*)}{\|\mathbf{x} - \bar{\mathbf{x}}\|_2 \cdot \|\mathbf{w}_k^* - \bar{\mathbf{w}}_k^*\|_2} \quad (10)$$

where  $\bar{\mathbf{x}}$  is a  $d$ -dimensional vector with all elements being the same value computed by averaging all elements in  $\mathbf{x}$ , as is likewise for  $\bar{\mathbf{w}}_k^*$ .

The above-defined similarity metric will be put into the softmax function as denoted by Equation (1) to obtain the final prediction. For cosine similarity and Pearson correlation metrics, the values output by the softmax may be extremely small, since  $\|f_k(\mathbf{x})\|_2 \leq 1$  in such situations. So, similar to (Gidaris and Komodakis 2018; Qi, Brown, and Lowe 2018), we multiply a learnable scale parameter  $s$ , i.e., putting  $s \cdot f_k(\mathbf{x})$  into the softmax for these two metrics.

**Optimization**

Similar to (Hariharan and Girshick 2017), we adopt a two-stage training procedure to train the proposed model.

**Stage 1:** At the first stage, we train the feature extractor  $\phi(\cdot)$  based on the base set  $\mathcal{D}_{base}$ . Given an image sample  $x_i$  with label  $y_i$ , we first compute its probability distribution  $\mathbf{p}_i = \{p_i^1, p_i^2, \dots, p_i^{k_{base}}\}$  with  $p_i^k = p(y = k | x_i)$ , and then define the cross-entropy loss as our objective function. To make the learned features easily generalize to the novel categories, we further introduce Squared Gradient Magnitude loss proposed in (Hariharan and Girshick 2017) to regularize representation learning. Thus, the objective function at

this stage can be defined as

$$\mathcal{L}_1 = \mathcal{L}_c + \lambda \mathcal{L}_s \quad (11)$$

where

$$\begin{aligned} \mathcal{L}_c &= -\frac{1}{N_{base}} \sum_{i=1}^{N_{base}} \sum_{k=1}^{K_{base}} \mathbf{1}(k = y_i) \log p_i^k \\ \mathcal{L}_s &= \frac{1}{N_{base}} \sum_{i=1}^{N_{base}} \sum_{k=1}^{K_{base}} (p_i^k - \mathbf{1}(k = y_i))^2 \|\mathbf{x}_i\|_2^2 \end{aligned} \quad (12)$$

where  $N_{base}$  is the number of training samples in all base categories,  $\mathbf{1}(k = y_i)$  is an indicator that equals 1 when  $k = y_i$  is true and 0 otherwise,  $\lambda$  is a parameter to balance two loss terms and it is set as 0.005 by following (Hariharan and Girshick 2017). At this stage, the model is trained using the SGD algorithm with a batch size of 256, momentum of 0.9 and weight decay of 0.0005. The learning rate is initialized as 0.1 and is divided by 10 for every 30 epochs.

**Stage 2:** We fix the parameters of the feature extractor and use both the base and novel set to train the other components, including  $\mathbf{w}^{init}$ , the parameters of GGNN and scale factor  $s$ . Similarly, we obtain its probability vector  $\mathbf{p}_i = \{p_i^1, p_i^2, \dots, p_i^K\}$  and use the cross-entropy loss as the objective function. To handle class imbalance, we sample the novel and base samples by 1:1 ratio in each batch. Besides, for inner product, we introduce an additional regularization term on the classifier weighs, thus the overall objective function can be defined as

$$\mathcal{L}_2 = -\frac{1}{N} \sum_{i=1}^N \sum_{k=1}^K \mathbf{1}(k = y_i) \log p_i^k + \eta \sum_{k=1}^K \|\mathbf{w}_k^*\|_2^2 \quad (13)$$

where  $N$  is the number of all training samples,  $\eta$  is a parameter to balance two loss terms and it is set as 0.001 empirically. In this stage, we train the model using SGD algorithm with a batch size of 1,000, momentum of 0.9, weight decay of 0.0001, and learning rate of 0.01.

## Experiment

### Graph Construction

The knowledge graph encodes the correlations among different categories. It can be constructed according to different prior knowledge. Here, we introduce two kinds of knowledge, i.e., semantic similarity and category hierarchy.

**Semantic similarity.** Semantic word of a specific category well carries its semantic information, and the semantic distance of two categories encodes their correlations. In other words, two categories are of high correlation if their semantic distance is small, and are of low correlation otherwise. Thus, we first exploit this property to construct the graph. Specifically, given two categories  $i$  and  $j$  with semantic words  $w_i$  and  $w_j$ , we first extract their semantic feature vector  $\mathbf{f}_i^w$  and  $\mathbf{f}_j^w$  using the GloVe model (Pennington, Socher, and Manning 2014) and compute their Euclidean distance  $d_{ij}$ . Then, we apply a monotone decreasing function  $a_{ij} = \lambda^{d_{ij} - \min_{d_{ik}, \forall k \neq i}}$  (we set  $\lambda = 0.4$ ) to map the distance to the correlation coefficient  $a_{ij}$ . Besides, we also consider the self-message of each node so that we set  $a_{ii} = 1$ .

**Category hierarchy.** Category hierarchy encodes category correlations via different levels of concept abstraction. Generally, the distance from one category to another indicates their correlation, where a small distance indicates a high correlation while a large distance indicates a low correlation. In this work, we also exploit this knowledge based on the WordNet (Miller 1995) to construct the graph. Concretely, given two categories  $i$  and  $j$ , we compute the shortest path from node  $i$  to  $j$  as the distance  $d_{ij}$  and we also apply a similar monotone decreasing function to map the distance to the correlation coefficient  $a_{ij}$ .

### Datasets

Unlike previous few-shot benchmark with low-resolution images and few novel categories, we consider the more realistic dataset with large scale base and novel categories: *ImageNet Few-Shot (ImageNet-FS)* dataset. To further verify our approach, we construct a more challenging dataset *ImageNet-6K*, which covers 5,000 novel categories.

**ImageNet-FS.** In this work, we first evaluate our proposed model on the widely used ImageNet-FS dataset. The dataset covers 1,000 categories from ILSVRC2012 and is divided into 389 base categories and 611 novel categories where 193 base categories and 300 novel categories are used for cross-validation and the remaining 196 categories and 311 novel categories are used for testing. Each base category has about 1,280 training images and 50 test images.

**ImageNet-6K.** To evaluate our proposed method on larger scale categories, we further construct a new benchmark ImageNet-6K that covers 6,000 categories. It contains 1,000 categories from ImageNet 2012 dataset with all the labeled training samples as the base categories, and we further select 5,000 categories from the ImageNet 2011 as novel categories. Concretely, 2,500 categories are used for validation and the rest 2,500 categories for final evaluation. Each category has 50 test samples.

For  $k$ -shot learning, only  $k$  labeled images of the novel categories are used. We follow previous works (Hariharan and Girshick 2017; Wang et al. 2018a) to repeat the process by 5 times and report the average accuracy. Here, we set  $k$  as 1, 2, 5, 10 on ImageNet-FS and ImageNet-6K.

**Evaluation Metrics.** We follow previous works (Wang et al. 2018a) to evaluate our proposed model on the top-5 accuracy of the novel categories (**Acc of novel**) and all (base + novel) categories (**Acc of all**).

### Comparison with State-of-the-Art

**Performance on ImageNet-FS dataset.** We present the results of the above metrics on 1, 2, 5, 10 shot learning on the ImageNet-FS in Table 1. As shown, our proposed model outperforms all existing methods by a sizable margin. Take the ‘‘novel’’ metric as example, our model achieves the accuracies of 62.1%, 70.9%, 78.4%, and 82.3%, outperforming current best results by 2.0%, 1.3%, 1.0%, and 0.3%, on 1, 2, 5, 10 shot learning, respectively.

Notably, compared with KTCH which also introduces word embedding as the external knowledge, we construct semantic correlation based on word embedding to transferring information between classifier weight and obtain an im-

Dataset	Method	Novel				All			
		1	2	5	10	1	2	5	10
ImageNet-FS	MN(Vinyals et al. 2016)	53.5	63.5	72.7	77.4	64.9	71.0	77.0	80.2
	PN(Snell, Swersky, and Zemel 2017)	49.6	64.0	74.4	78.1	61.4	71.4	78.0	80.0
	SGM(Hariharan and Girshick 2017)	54.3	67.0	77.4	81.9	60.7	71.6	80.2	<b>83.6</b>
	SGM w/ G(Hariharan and Girshick 2017)	52.9	64.9	77.3	82.0	63.9	71.9	80.2	<b>83.6</b>
	AWG(Gidaris and Komodakis 2018)	53.9	65.5	75.9	80.3	65.1	72.3	79.1	82.1
	PMN(Wang et al. 2018a)	53.3	65.2	75.9	80.1	64.8	72.1	78.8	81.7
	PMN w/ G(Wang et al. 2018a)	54.7	66.8	77.4	81.4	65.7	73.5	80.2	82.8
	LSD(Douze et al. 2018)	57.7	66.9	73.8	77.6	-	-	-	-
	KTCH(Li et al. 2019)	58.1	67.3	77.6	81.8	-	-	-	-
	IDeMe-Net(Chen et al. 2019c)	60.1	69.6	77.4	80.2	-	-	-	-
	Ours(CosSim)	61.4	70.4	78.4	82.2	67.7	74.7	<b>80.9</b>	<b>83.6</b>
	Ours(PearsonCorr)	61.5	70.6	<b>78.5</b>	<b>82.3</b>	67.5	74.4	80.7	83.5
Ours(InnerProduct)	<b>62.1</b>	<b>70.9</b>	78.4	<b>82.3</b>	<b>68.3</b>	<b>75.2</b>	80.8	83.5	
ImageNet-6K	SGM(Hariharan and Girshick 2017)	26.7	35.0	44.7	51.5	37.7	44.8	53.1	58.5
	SGM w/ G(Hariharan and Girshick 2017)	26.3	35.5	<b>46.2</b>	52.0	39.4	46.4	<b>54.4</b>	<b>58.8</b>
	AWG(Gidaris and Komodakis 2018)	27.6	35.9	45.0	49.4	39.3	45.2	52.1	55.4
	Ours(CosSim)	<b>30.5</b>	<b>37.5</b>	46.0	51.7	41.0	46.8	53.8	58.5
	Ours(PearsonCorr)	30.4	37.2	45.9	51.8	<b>41.1</b>	46.8	53.7	58.5
	Ours(InnerProduct)	29.5	36.8	46.1	<b>52.1</b>	40.8	<b>46.9</b>	<b>54.4</b>	58.7

Table 1: Top5-accuracy in “novel” and “all” metrics of our model and current state-of-the-art methods on the ImageNet-FS and ImageNet-6K datasets. For fair comparison, all the methods use ResNet-50 for feature extraction. Some methods train an additional generator to hallucinate extra training samples for the novel categories (w/ G). The best and second best results are highlighted in **bold** and *italic*, respectively. “-” denotes the corresponding result is not provided.

provement of 4.0%, 3.6% in 1-shot and 2-shot setting, which shows our superiority of utilizing external knowledge. Besides, it is worth noting that the improvement is more notable if the samples of novel categories are fewer, e.g., 7.4% on 1-shot learning v.s. 0.3% on 10-shot learning. One possible reason for this phenomenon is that learning from fewer samples is more challenging and thus depends more on prior knowledge. Similar improvements are observed on the “all” metrics.

**Performance on ImageNet-6K dataset.** All the foregoing comparisons focus on recognizing about 500 categories. In fact, there are much more categories in a real-world setting. To evaluate the proposed model on larger scale categories, we further conduct experiments on the ImageNet-6K dataset. As SGM (Hariharan and Girshick 2017) and AWG (Gidaris and Komodakis 2018) are current leading methods, we use the released codes to implement these methods on this dataset for comparison. Following (Hariharan and Girshick 2017), we train models with and without generating samples for the novel categories. The comparison results are presented in Table 1. Even though ImageNet-6K is more challenging and much larger in category size, our proposed model still outperforms existing methods. Specifically, it obtains the “novel” accuracy of 30.5%, 37.5%, an improvement of 2.9%, 1.6% on the 1-shot and 2-shot learning compared with existing methods. This comparison suggests the proposed method is scalable in category size.

## Ablative study

**Analysis of Knowledge Embedding.** The core of our proposed model is the prior knowledge graph that correlates base and novel categories to help regularize parameter propagation. As discussed above, our model follows SGM (Har-

iharan and Girshick 2017) to use ResNet-50 as feature extractor and also use identical loss function for optimization, thus SGM can be regarded as the baseline of our model. In this part, we emphasize the comparison with SGM to evaluate the effectiveness of knowledge embedding model. As shown in Table 1, our framework significantly outperforms SGM, with accuracy improvement of 7.8%, 7.6% on two metrics in 1-shot setting.

To further analyze the effect of knowledge embedding, we further replace the category correlations with other non-informative form to demonstrate its benefit. Specifically, we consider the following two variants: 1) Uniform graph in which all the correlation values are uniformly set as  $\frac{1}{K}$  and 2) Random graph in which all the correlation values are randomly selected from a uniform distribution. The comparison results are presented in Table 2. We find that these two variants perform comparably with each other as both incur no additional information. Note that they achieve slightly better results than the baseline SGM. One possible reason is that knowledge propagation can help to better learn the classifier weights of novel categories. Still, our model with prior knowledge embedding significantly outperforms both two variants on all metrics for 1, 2, 5, 10 shot learning. These comparisons clearly demonstrate the benefit of knowledge embedding.

**Analysis on Different Kinds of Knowledge.** The correlations between categories can be constructed based on different kinds of knowledge, e.g., semantic similarity and category hierarchy in the paper. Here, we further conduct experiments with these two kinds of knowledge and present their results in Table 2. We find that introducing both kinds of knowledge leads to obvious improvement than the baseline SGM and those with non-informative correlations, which

Graph	Novel				All			
	1	2	5	10	1	2	5	10
u-graph	53.4	67.4	77.8	81.5	63.8	73.3	80.3	82.9
r-graph	54.4	67.4	77.8	81.9	64.5	73.3	80.5	83.2
c-graph	60.1	69.4	78.1	82.1	67.0	74.4	80.7	83.3
s-graph	<b>62.1</b>	<b>70.9</b>	<b>78.4</b>	<b>82.3</b>	<b>68.3</b>	<b>75.2</b>	<b>80.8</b>	<b>83.5</b>

Table 2: Top5-accuracy in “novel” and “all” metrics of our proposed model with semantic correlation knowledge (s-graph), with category hierarchy knowledge (c-graph), uniform correlation value (u-graph), and random correlation value (r-graph).

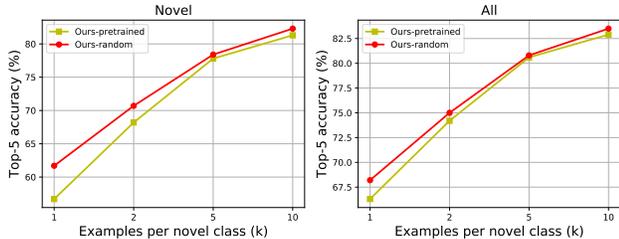


Figure 4: Top5-accuracy in “novel” and “all” metrics of our model that initializes  $\mathbf{W}_{base}$  with random values (Ours-random) and with the parameters pre-trained at the first training stage (Ours-pretrained).

suggests that our model can adapt to different kinds of knowledge. On the other hand, we find introducing semantic similarity knowledge achieves slightly better performance than introducing category hierarchy. One possible reason is that semantic similarity provides stronger and more direct correlations among different categories.

**Analysis on Different Similarity Metrics.** In this work, the classifier is viewed as similarity measure between the input feature and the classifier weight, and three different similarity metrics, including inner product, cosine similarity and Pearson correlation coefficient, are evaluated. As shown in Table 1, with different metrics, we still achieve better performance than the previous best result. Besides, we notice that inner product leads to slightly better accuracy than other metrics on ImageNet-FS, while cosine similarity and Pearson correlation coefficient perform better in much larger scale scenarios, for example, cosine similarity achieves higher accuracy than the inner product, with an increase of 1.0% and 0.7% for 1-shot and 2-shot setting in ImageNet-6K.

**Analysis on Classifier Weight Initialization.** At the second stage of the training process, we randomly initialize both  $\mathbf{W}_{base} = \{\mathbf{w}_k\}_{k=1}^{K_{base}}$  and  $\mathbf{W}_{novel} = \{\mathbf{w}_k\}_{k=K_{base}+1}^K$  to initialize the graph nodes. Actually, it is also intuitive to initialize  $\mathbf{W}_{base}$  with the parameters pre-trained at the first stage and initialize  $\mathbf{W}_{novel}$  randomly. Here, we further compare these two initialization settings and present the results in Figure 4. We find that random initialization leads to much better results, up to 5.0% accuracy improvement on 1-shot learning for the novel categories. The reason for this phe-

Method	Novel				
	1	2	3	4	5
<b>ImageNet2012/2010</b>					
NN	34.2	43.6	48.7	52.3	54.0
SGM(Hariharan and Girshick 2017)	31.6	42.5	49.0	53.5	56.8
PPA(Qiao et al. 2018b)	33.0	43.1	48.5	52.5	55.4
LSD(Douze et al. 2018)	33.2	44.7	50.2	53.4	57.6
KTCH(Li et al. 2019)	39.0	48.9	54.9	<b>58.7</b>	60.5
Ours	<b>42.5</b>	<b>50.3</b>	<b>55.4</b>	58.4	<b>60.7</b>

Table 3: Top5-accuracy on ImageNet2012/2010 dataset.

nomenon is updating  $\mathbf{W}_{base}$  from scratch enables training the graph propagation mechanism using sufficient samples of the base categories and can be easily generalized to update the weights of the novel categories.

## Further evaluation

**Comparison on the large scale few-shot benchmark ImageNet2012/2010.** Here we also evaluate our method on another large scale few-shot benchmark **ImageNet2012/2010**, proposed by (Li et al. 2019). Briefly, in ImageNet2012/2010, all 1,000 categories from ILSVRC2012 are considered as the base categories, and 360 novel categories from ILSVRC2010, which are not overlapped with the base categories, are used as the novel categories. The base categories cover 200,000 labeled samples and the novel categories have 150 test samples. Same with (Li et al. 2019), We set  $k$  as 1, 2, 3, 4, 5. Since we are not able to use the same training samples as (Li et al. 2019) did, we randomly select training samples for 5 times as we did on ImageNet-FS dataset and report the mean accuracy as the final result.

We compare with several methods on this benchmark including NN (nearest neighbor) (Li et al. 2019), SGM (Hariharan and Girshick 2017), PPA (parameter prediction from activations) (Qiao et al. 2018a), LSD (large-scale diffusion) (Douze et al. 2018), and KTCH (Li et al. 2019). The comparative results are shown in Table 3. It can be seen that our method achieves superior performance in most settings and exceeds over the best reported results by 3.5%, 1.4%, 0.5% in 1, 2, 3 shot. Notably, KTCH also incorporates semantic word embedding as prior knowledge. Both KTCH and our method achieve significant improvements over other competing methods. Moreover, our method obtains superior results than KTCH, demonstrating its effectiveness.

## Conclusion

In this work, we explore incorporating prior knowledge of category correlations to guide exploiting knowledge from base categories to help learn the classifier weights of novel categories. To this end, we formulate a novel Knowledge Graph Transfer Network model, which correlates classifier weights with a graph constructed based on their correlations and introduce a graph neural network to iteratively propagate and update classifier weights. Extensive experiments on the widely used ImageNet-FS and newly constructed ImageNet-6K dataset demonstrate the effectiveness of our proposed model over state-of-the-art methods.

## References

- [Chen et al. 2018a] Chen, T.; Chen, R.; Nie, L.; Luo, X.; Liu, X.; and Lin, L. 2018a. Neural task planning with and-or graph representations. *IEEE Transactions on Multimedia* 21(4):1022–1034.
- [Chen et al. 2018b] Chen, T.; Lin, L.; Chen, R.; Wu, Y.; and Luo, X. 2018b. Knowledge-embedded representation learning for fine-grained image recognition. *arXiv preprint arXiv:1807.00505*.
- [Chen et al. 2018c] Chen, T.; Wu, W.; Gao, Y.; Dong, L.; Luo, X.; and Lin, L. 2018c. Fine-grained representation learning and recognition by exploiting hierarchical semantic embedding. *arXiv preprint arXiv:1808.04505*.
- [Chen et al. 2019a] Chen, T.; Xu, M.; Hui, X.; Wu, H.; and Lin, L. 2019a. Learning semantic-specific graph representation for multi-label image recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, 522–531.
- [Chen et al. 2019b] Chen, T.; Yu, W.; Chen, R.; and Lin, L. 2019b. Knowledge-embedded routing network for scene graph generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6163–6171.
- [Chen et al. 2019c] Chen, Z.; Fu, Y.; Wang, Y.-X.; Ma, L.; Liu, W.; and Hebert, M. 2019c. Image deformation meta-networks for one-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8680–8689.
- [Douze et al. 2018] Douze, M.; Szlam, A.; Hariharan, B.; and Jégou, H. 2018. Low-shot learning with large-scale diffusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3349–3358.
- [Finn, Abbeel, and Levine 2017] Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 1126–1135. JMLR. org.
- [Garcia and Bruna 2017] Garcia, V., and Bruna, J. 2017. Few-shot learning with graph neural networks. *arXiv preprint arXiv:1711.04043*.
- [Gidaris and Komodakis 2018] Gidaris, S., and Komodakis, N. 2018. Dynamic few-shot visual learning without forgetting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4367–4375.
- [Hariharan and Girshick 2017] Hariharan, B., and Girshick, R. 2017. Low-shot visual recognition by shrinking and hallucinating features. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 3037–3046. IEEE.
- [He et al. 2016] He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.
- [Hui, Chen, and Chen 2019] Hui, X.; Chen, R.; and Chen, T. 2019. Graph attention propagation for few-shot learning. In *Proceedings of the ACM Turing Celebration Conference-China*, 103. ACM.
- [Jiang et al. 2018] Jiang, C.; Xu, H.; Liang, X.; and Lin, L. 2018. Hybrid knowledge routed modules for large-scale object detection. In *Advances in Neural Information Processing Systems*, 1552–1563.
- [Kim et al. 2019] Kim, J.; Oh, T.-H.; Lee, S.; Pan, F.; and Kweon, I. S. 2019. Variational prototyping-encoder: One-shot learning with prototypical images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9462–9470.
- [Kipf and Welling 2016] Kipf, T. N., and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.
- [Krizhevsky, Sutskever, and Hinton 2012] Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.
- [Lee et al. 2018] Lee, C.-W.; Fang, W.; Yeh, C.-K.; and Frank Wang, Y.-C. 2018. Multi-label zero-shot learning with structured knowledge graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1576–1585.
- [Li et al. 2017] Li, Z.; Zhou, F.; Chen, F.; and Li, H. 2017. Meta-sgd: Learning to learn quickly for few-shot learning. *arXiv preprint arXiv:1707.09835*.
- [Li et al. 2019] Li, A.; Luo, T.; Lu, Z.; Xiang, T.; and Wang, L. 2019. Large-scale few-shot learning: Knowledge transfer with class hierarchy. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7212–7220.
- [Marino, Salakhutdinov, and Gupta 2017] Marino, K.; Salakhutdinov, R.; and Gupta, A. 2017. The more you know: Using knowledge graphs for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2673–2681.
- [Miller 1995] Miller, G. A. 1995. Wordnet: a lexical database for english. *Communications of the ACM* 38(11):39–41.
- [Pennington, Socher, and Manning 2014] Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.
- [Qi, Brown, and Lowe 2018] Qi, H.; Brown, M.; and Lowe, D. G. 2018. Low-shot learning with imprinted weights. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5822–5830.
- [Qiao et al. 2018a] Qiao, S.; Liu, C.; Shen, W.; and Yuille, A. 2018a. Few-shot image recognition by predicting parameters from activations. In *IEEE Conference on Computer Vision and Pattern Recognition*.
- [Qiao et al. 2018b] Qiao, S.; Liu, C.; Shen, W.; and Yuille, A. L. 2018b. Few-shot image recognition by predicting parameters from activations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7229–7238.
- [Snell, Swersky, and Zemel 2017] Snell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, 4077–4087.
- [Sung et al. 2018] Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P. H.; and Hospedales, T. M. 2018. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1199–1208.
- [Vinyals et al. 2016] Vinyals, O.; Blundell, C.; Lillicrap, T.; Wierstra, D.; et al. 2016. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, 3630–3638.
- [Wang et al. 2018a] Wang, Y.-X.; Girshick, R.; Hebert, M.; and Hariharan, B. 2018a. Low-shot learning from imaginary data. In *IEEE Conference on Computer Vision and Pattern Recognition*.
- [Wang et al. 2018b] Wang, Z.; Chen, T.; Ren, J.; Yu, W.; Cheng, H.; and Lin, L. 2018b. Deep reasoning with knowledge graph for social relationship understanding. *arXiv preprint arXiv:1807.00504*.
- [Wang, Ye, and Gupta 2018] Wang, X.; Ye, Y.; and Gupta, A. 2018. Zero-shot recognition via semantic embeddings and knowledge graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6857–6866.

[Yang et al. 2019] Yang, W.; Wang, X.; Farhadi, A.; Gupta, A.; and Mottaghi, R. 2019. Visual semantic navigation using scene priors. In *International Conference on Machine Learning*.

[Zhu, Mumford, and others 2007] Zhu, S.-C.; Mumford, D.; et al. 2007. A stochastic grammar of images. *Foundations and Trends® in Computer Graphics and Vision* 2(4):259–362.