Large-Scale Few-Shot Learning: Knowledge Transfer With Class Hierarchy

Aoxue Li[1*] Tiange Luo[1*] Zhiwu Lu[2] Tao Xiang[3] Liwei Wang[1]
[1]School of EECS, Peking University, Beijing 100871, China
[2]School of Information, Renmin University of China, Beijing 100872, China
[3]Department of Electrical and Electronic Engineering, University of Surrey, UK
{lax,luotg,wanglw}@pku.edu.cn, luzhiwu@ruc.edu.cn, t.xiang@surrey.ac.uk

Abstract

Recently, large-scale few-shot learning (FSL) becomes topical. It is discovered that, for a large-scale FSL problem with 1,000 classes in the source domain, a strong baseline emerges, that is, simply training a deep feature embedding model using the aggregated source classes and performing nearest neighbor (NN) search using the learned features on the target classes. The state-of-the-art large-scale FSL methods struggle to beat this baseline, indicating intrinsic limitations on scalability. To overcome the challenge, we propose a novel large-scale FSL model by learning transferable visual features with the class hierarchy which encodes the semantic relations between source and target classes. Extensive experiments show that the proposed model significantly outperforms not only the NN baseline but also the state-of-the-art alternatives. Furthermore, we show that the proposed model can be easily extended to the large-scale zero-shot learning (ZSL) problem and also achieves the state-of-the-art results.

1. Introduction

In the past five years, the object recognition research has focused on large-scale recognition problems such as the ImageNet ILSVRC challenges [22]. Deep convolutional neural network (DCNN) based models [24, 8] have achieved super-human performance on the ILSVRC 1K recognition task. However, most existing object recognition models, particularly those DCNN based ones, require hundreds of image samples to be collected for each object class; many of the object classes are rare and it is very hard to collect sufficient training samples, even with social media. Therefore, it is highly desirable to develop object recognition models that require only few training samples/shots per object class.

To overcome this challenge, meta-learning based few-shot learning (FSL) [28, 21, 7, 20, 19] has become topical.

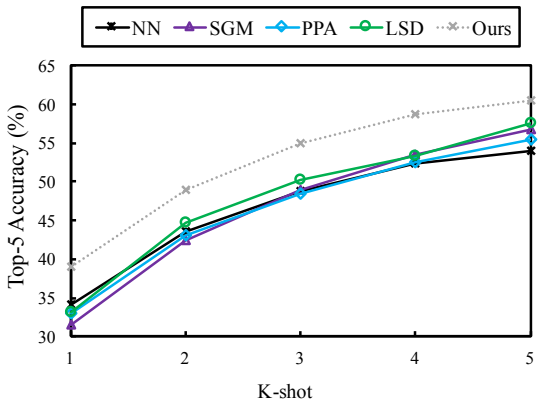

Figure 1. Comparative results for large-scale FSL on the ImNet dataset [11]. The top-5 accuracy over target class samples is used as the evaluation metric. It can be observed that the state-of-the-art large-scale FSL methods struggle to beat the simple NN baseline, suggesting intrinsic limitations on scalability. Notations: NN – nearest neighbor (NN) search performed in a learned feature space using K samples per target class as the references; SGM – FSL with the squared gradient magnitude (SGM) loss [7]; PPA – parameter prediction from activations (PPA) [20]; LSD – large-scale diffusion (LSD) [2]; Ours – the proposed model.

FSL is inspired by the fact that human can easily recognize novel objects with a few samples thanks to the ability to knowledge transfer. Similarly, in the FSL problem, we are provided with a set of source classes and a set of target classes under the setting that: (1) The target classes have no overlap with the source classes in the label space; (2) Each source class has sufficient labeled samples, whereas each target class has only a few labeled samples. FSL thus aims to transfer knowledge from the source to target classes.

The focus of this work is on the large-scale FSL setting with a large number of source classes provided. This is very different from the most widely used meta-learning benchmarks such as mini-ImageNet [29] which contains 64 source classes with 600 samples in each class. Yet it is more realistic – after all, we have 1,000s of classes in ImageNet

*Equal contribution.

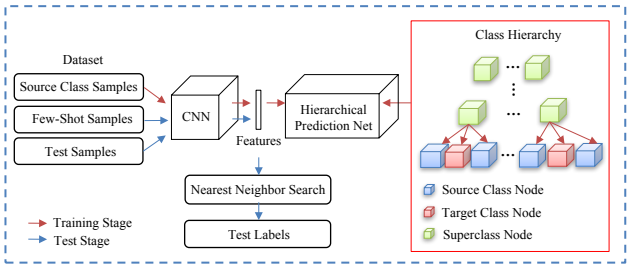

Figure 2. Overview of the proposed model.

that we can use, so why not include more source classes when it comes to FSL? It is noted that a deep feature embedding model learned to classify a large number of source classes would yield a transferable feature representation that can be directly applied to a wide range of vision problems [15]. This suggests a strong baseline for the large-scale FSL setting, that is, learning the feature embedding using all source classes together, followed by simply using the embedding model to extract deep features for target class samples for nearest neighbor search based classification. Indeed, when several latest large-scale FSL models [7, 20, 2] are evaluated on the large-scale ILSVRC2012/2010 (ImNet) dataset [11], they all struggle to beat this forgotten baseline. In this dataset, 1,000 classes of ILSVRC2012 are used as source classes and 360 classes of ILSVRC2010 (not included in ILSVRC2012) are used as target classes. Each target class is provided with K (≤ 5) labeled samples. The feature embedding is obtained by training a ResNet50 [8] for classifying the ImageNet ILSVRC2012 1K classes. The comparative results in Figure 1 show that the simple nearest neighbor (NN) search baseline with the learned feature embedding is competitive when compared against the state-of-the-art FSL models[1]. This suggests that existing FSL models are intrinsically limited on scalability. It also implies that most of knowledge transfer in SGM, LSD and PPA is achieved by the transferable features extracted by the deep feature embedding model. This observation indicates that learning more transferable visual features, which better represent target class samples, is an alternative way to tackle the large-scale FSL.

In this work, we thus propose a novel FSL model by exploiting a class hierarchy shared by both source and target classes to learn a more transferable feature embedding model. Our idea is simple: the semantic relations between source classes and target classes are used as the prior knowledge to help learn a more transferable feature embedding for recognizing the target class samples. In our work, the semantic relations are explicitly encoded to a tree-like class hierarchy by a data-driven approach based on a public text corpus, without the need of a human-annotated taxonomy. Such a tree can thus easily cover all object classes ever existed. Even though the source and target classes do not overlap at the bottom (leaf) layer of a class hierarchy, they share (superclass) labels in the top layers. Specifically, in such class hierarchy, both target classes and source classes are represented as the leaves (i.e., class nodes); semantic similar classes (including both source and target classes) are grouped, and each cluster then forms a parent node (i.e., a superclass node) in the upper layer of the tree (see the **red** box in Figure 2). To integrate the prior knowledge from the class hierarchy, we propose a novel hierarchical prediction net which explicitly encodes the class hierarchy into the classification procedure. During the training stage (see the **red** arrow in Figure 2), the source class samples only are fed into a convolutional neural network (CNN) followed by the proposed hierarchical prediction net. Since the source classes and target classes are certain to have some common superclasses, our hierarchical prediction net enables us to learn transferable features for FSL on the target classes. During the test stage (see the **blue** arrow in Figure 2), we extract visual features of test samples and the few training samples (both from target classes) using the proposed feature learning model. The test samples are then recognized by a simple nearest neighbor search using the visual features of few-shot samples (from target classes) as the references. Furthermore, our feature learning model can be easily extended to the closely related zero-shot learning (ZSL) problem [32, 17, 29, 31, 14]. Experimental results on several benchmark datasets show that our model achieves state-of-the-art results on both large-scale FSL and ZSL problems (see Figure 1 and Tables 1&4&5).

Our contributions are three-fold: 1) We make an important observation that there exists a strong baseline based on deep feature embedding over source classes and nearest neighbor search over target classes. Existing large-scale FSL models' advantage over this baseline diminishes. 2) We propose a novel large-scale FSL model that can learn transferable visual features by exploiting a class hierarchy which encodes the semantic relations between source and target classes. Extensive experiments show that the proposed model outperforms not only the NN baseline but also the state-of-the-art alternatives. 3) We show that the proposed model can be easily extended to zero-shot learning (ZSL) problem [29, 17, 32, 11], with state-of-the-art results obtained. This further validates the effectiveness of knowledge transfer with class hierarchy.

2. Related Work

2.1. Few-Shot Learning

Learning to learn [27] is topical in the machine learning community, and one of its well-received applications is few-shot learning (FSL). Meta-learning based approaches

[1]The results of PPA, LSD and SGM are obtained by training the original code provided in their papers using the large-scale ImNet dataset.

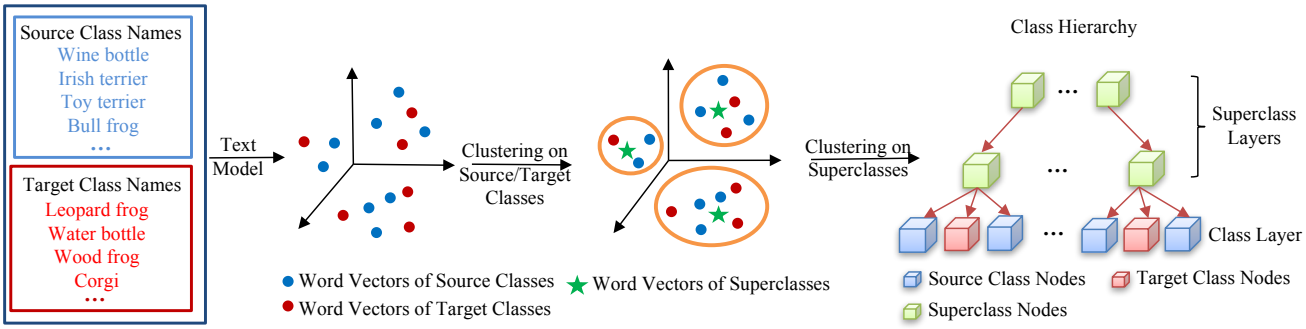

Figure 3. Illustration of tree-structured class hierarchy construction by a data-driven approach, without the need of a human-annotated taxonomy. Note that no image samples are used here.

[17, 21, 28, 10] have dominated. Apart from metric learning solutions [25, 28], another promising approach is learning to optimize [21, 3]. More recently, methods based on feature hallucination and synthesis [7, 23] or predicting parameters of the network [20, 19] have been developed. Most of these works consider small datasets like Omniglot, CIFAR, or miniImageNet. The more recent works [7, 2, 20, 30] start to pay more attention to large scale FSL. However, the strong baseline mentioned earlier has never been considered as a serious competitor. In this work, we only focus on FSL under large-scale. The proposed FSL model is completely different from the existing FSL approaches in that the class hierarchy is exploited for learning more transferable features. We demonstrate that our FSL model achieves state-of-the-art results on large-scale datasets and it is the only one that can consistently beat the strong baseline (see Figure 1 and Tables 1&4).

2.2. Zero-Shot Learning

As a closely related problem, ZSL assumes that no visual samples are given for target classes. To address such data sparsity issue, existing ZSL approaches employ attributes or textual description of object classes as an intermediate feature space that can transfer knowledge across source/target classes [26, 12, 32, 17, 29, 31, 14]. This is made possible by directly learning a mapping from the visual image space to a semantic space (e.g., attribute space) only with the source class data. However, when the learned mapping is applied to the target class data, the domain gap occurs [11, 33], which is known as the biggest challenge in ZSL. In this work, our model is seamlessly extended to ZSL, because we only exploit the source class data, along with the semantic relations between target classes and source classes. Since our model enables us to learn transferable visual features for both source and target class samples, the domain gap issue can be alleviated. We show that our transferable features can improve the performance of the mapping-learning-based ZSL model (see Table 5).

2.3. Knowledge Transfer with Class Hierarchy

In the area of FSL/ZSL, little attention has been paid to knowledge transfer with the class hierarchy. Two exceptions are: (1) The relation between attributes and superclasses (from the class hierarchy) is learned for semantic embedding in ZSL [9]; (2) The class hierarchy is used to define a semantic space for ZSL [1]. However, in these two works, the class hierarchy is obtained from the manually-defined hierarchical taxonomy, which needs the additional cost to collect. In this work, our model is more scalable by generating the class hierarchy automatically with data-driven clustering overall source/target classes. In addition, the class hierarchy is not involved in feature learning in [1, 9].

3. Model

3.1. Problem Definition

We first formally define the large-scale FSL problem as follows. Let S_{source} denote the set of source classes and S_{target} denote the set of target classes. These two sets of classes are disjoint, i.e., $S_{source} \cap S_{target} = \phi$. We are given a large-scale sample set D_{source} from source classes S_{source}, a few-shot sample set D_{target} from target classes S_{target}, and a test set D_{test} from target classes S_{target}. D_{source} contains sufficient labeled samples for each source class, whereas D_{target} contains only a few (≤ 5 in this paper) labeled samples for each target class. The goal of large-scale FSL is to obtain good classification results on D_{test}. Our approach to large-scale FSL consists of two phases: transferable visual feature learning and label inference with learned features. The details of these two phases are given below.

3.2. Feature Learning

We propose a novel transferable feature learning model for large-scale FSL. In this model, a tree-structured class hierarchy is first constructed to encode semantic relations

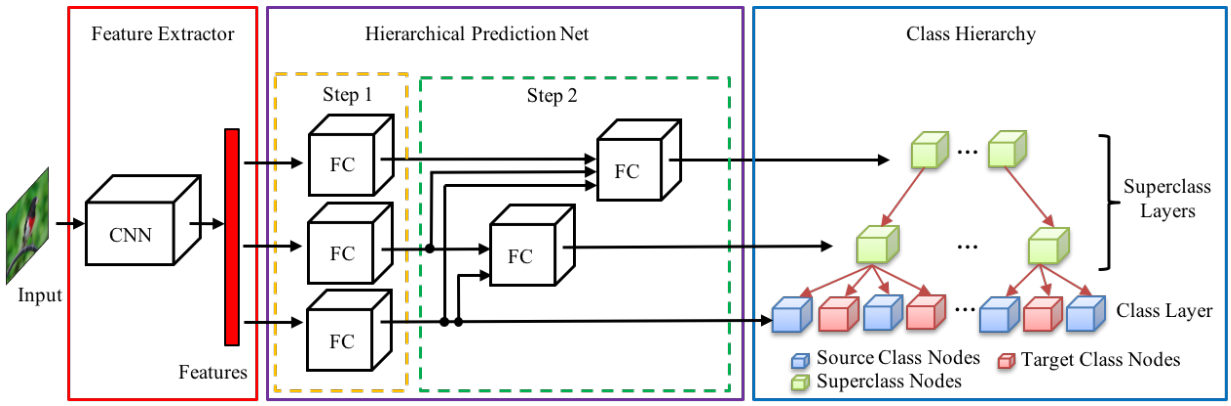

Figure 4. Overview of the proposed feature learning model. In this model, a tree-structured class hierarchy is first constructed to encode semantic relations between source and target classes; after that, by proposing a novel hierarchical prediction net, we integrate the prior knowledge from the class hierarchy to learn transferable features. Notation: 'FC'– fully-connected network.

between source and target classes; after that, a hierarchical prediction net is proposed to integrate the prior knowledge from the class hierarchy for learning transferrable visual features for large-scale FSL. The proposed model is trained with the source class data D_{source} as well as the semantic relations between source and target classes.

We first provide a data-driven approach to generate tree-structured class hierarchy, without the need of a human-annotated taxonomy. Specifically, we first represent each source/target class name using a word vector extracted from a skip-gram text model [16] trained on a corpus of 4.6M Wikipedia documents. In our class hierarchy, both target classes and source classes are exploited as leaves (i.e., class nodes) of the tree. They form the bottom class layer of the class hierarchy. Starting from the leaves, we obtain the nodes in the upper layer by clustering over the word vectors of the nodes in the lower layer. Each cluster then forms a parent node (i.e., a superclass node) in the upper layer of the tree, and the word vector of the superclass is an average of word vectors of its children classes. Superclass nodes in the same layer form a superclass layer. By using such an approach, we can obtain a tree-structured class hierarchy, which consists of n superclasses layers and one class layer (see Figure 3). To simplify, we denote l_1 for the class layer and $l_i (i = 2, ..., n + 1)$ for n superclass layers. Since the superclasses share across both source and target classes, our model is expected to represent well the target class samples.

With the obtained class hierarchy, we now propose the transferable feature learning model. Concretely, we extend a CNN model by a hierarchical prediction net, as shown in the **purple** box in Figure 4. This net consists of two steps for predicting the superclass labels, using the shared CNN generated features (see the **red** box in Figure 4). The first step is to predict the labels at different class/superclass layers (see the **yellow** dashed box in Figure 4), so that the superclasses shared across source classes and target classes make the learned features suitable for representing the target classes. The second step is to encode the hierarchical structure of class/superclass layers into superclass label prediction (see the **green** dashed box in Figure 4). That is, we infer superclass labels of each layer by combining the prediction results of the same and lower class/superclass layers obtained by the first step. Since the hierarchy between consecutive layers is shared and transferred across the source and target classes, the encoded hierarchical structure thus further improves the transferability of the learned deep visual features. The technical details are given as follows.

For the first class/superclass label prediction step, we add $n+1$ unshared fully-connected (FC) networks with softmax layers on top of the CNN model (see the **yellow** dashed box in Figure 4). Given an object sample, each FC network thus predicts the probability distribution of classes/superclasses at the corresponding layer.

For the second superclass label prediction step, we encode the hierarchical structure among class/superclass layers by using n unshared FC networks, each inferring the labels at the corresponding superclass layer (see the **green** dashed box in Figure 4). To model the hierarchical structure between consecutive layers, we concatenate all the outputs of the current layer and its lower layers in the first step as the input and feed it into the FC network in the second step at the corresponding layer. To be specific, for the FC network corresponding to the lowest superclass layer (i.e., layer l_2), we combine the outputs of the bottom two layers in the first prediction step as its input. The output of this FC network is the final prediction results concerning the lowest superclass layer. Its formal formulation is given as:

$$\hat{p}_{l_2} = \mathcal{F}^2_{l_2}(p_{l_1} \oplus p_{l_2}) \tag{1}$$

where p_{l_1} denotes the output of the bottom FC network in the first prediction step which means the prediction results of class layer, p_{l_2} denotes the output of second FC network

in the first step which means the prediction results of the lowest superclass layer in the first step. $\oplus(\cdot)$ is a concatenation operator by channel, and $\mathcal{F}^2_{l_2}(\cdot)$ is a forward step of the FC network corresponding to the layer l_2 in the second prediction step. The output $\hat{p}_{l_2}$ denotes a final predicted distribution over all possible superclass labels at the second layer of the hierarchy.

Similarly, we can also infer the superclass labels in the $l_i (i = 3, ..., n+1)$ layer by an FC network with the prediction results of the layers $\{l_j : j <= i\}$ in the first prediction step as its input. Therefore, by merging these hierarchical inference steps with the original class label prediction step, we define the loss function for an image x as follows:

$$\begin{aligned} p_{l_i} &= \mathcal{F}^1_{l_i}(G(x)),\ i = 1, ..., n+1 \\ \hat{p}_{l_i} &= \mathcal{F}^2_{l_i}(\mathop{\oplus}_{j=1}^{i} p_{l_j}),\ i = 2, ..., n+1 \\ \mathcal{L}(x, Y; \Theta) &= \mathcal{L}_{cls}(y_{l_1}, p_{l_1}) + \sum_{i=2}^{n+1} \lambda_i \mathcal{L}_{cls}(y_{l_i}, \hat{p}_{l_i}) \end{aligned} \tag{2}$$

where G denote a forward step of the CNN for feature extraction. $\mathcal{F}^1_{l_i}$ and $\mathcal{F}^2_{l_i}$ respectively denote a forward step of the FC network corresponding to layer l_i in the first and second step. p_{l_i} denotes the predicted distribution over possible classes/superclasses in layer l_i in the first step. $\hat{p}_{l_i}$ denotes the final predicted distribution over possible superclasses in layer l_i. $\oplus$ is a concatenation operator by channel. $Y = \{y_{l_i}, i = 1, .., n+1\}$ collects the true class/superclass labels of the image x, where y_{l_i} denotes the label corresponding to layer l_i. Θ denotes the parameters of the full network. $\mathcal{L}_{cls}$ denotes the cross entropy loss for classification, and λ_i weights these losses.

3.3. Label Inference

Once our feature learning model is trained with the source class data, it can be used to extract features for image samples from target classes (i.e., samples from D_{target} and D_{test}). With these visual features, we directly use a simple nearest neighbor search method for inferring the labels of test samples from D_{test}. Specifically, for each target class, we compute the average of the visual features of its few-shot samples as its reference. Given a test image, we compute its cosine distance to each class reference and predict its class label as the one with the smallest distance.

3.4. Extension to Large-Scale ZSL

Although the proposed feature learning model is originally designed for large-scale FSL, it can be easily extended to large-scale ZSL: the training data contain all the instances from the large-scale source(seen) classes but without any visual samples from target(unseen) classes. Concretely, we first construct a tree-structured class hierarchy using word vectors of all seen and unseen class names as in Sec. 3.2. With the obtained class hierarchy, we further train our deep feature learning model over the whole training set (i.e., all seen class samples) as before. When the visual features are extracted with our feature learning model, we can infer the labels of test images using the state-of-the-art mapping-learning-based ZSL model [11]. Experimental results on a large-scale benchmark dataset show that our method can improve the representative model [11] and create a new state-of-the-art for large-scale ZSL.

4. Experiments and Discussion

4.1. Large-Scale FSL

4.1.1 Experimental Setup

Dataset and Settings. A new benchmark dataset is derived from ILSVRC2012/2010 (ImNet) [11] for performance evaluation. This dataset is organized into three parts: a training set of many labeled source class samples, a few-shot set of few labeled target class samples, and a test set of the rest target class samples. Concretely, the **1,000** classes of ILSVRC2012 are used as the source classes, and the 360 classes of ILSVRC2010 (not included in ILSVRC2012) are used as the target classes, as in [11]. A strong baseline is obtained directly by using ImageNet ILSVRC2012 1K classes pretrained ResNet50 [8] as the deep feature embedding model, followed by simply performing nearest neighbor search based classification over target class samples. We compare our model with four large-scale FSL models: (1) NN – the nearest neighbor (NN) search-based strong baseline performed in the feature space using K samples per target class as the references. The knowledge transfer is realized with the feature space formed with pretrained ResNet50. (2) SGM – the low-shot learning model using the squared gradient magnitude (SGM) loss [7]. (3) PPA – the parameter prediction with activations (PPA) model [20]. (4) LSD – the low-shot learning model with large-scale diffusion (LSD). The top-5 accuracy overall test images from 360 target classes is computed for each K (i.e. the number of samples per target class) as the evaluation metric.

Network Architecture and Training Details. In this work, we construct a 3-superclass-layer class hierarchy for transferable feature learning. These three superclass layers have 200, 40, and 8 superclasses, respectively. The clustering method used for constructing class hierarchy is k-means clustering over word vectors of source/target class names. In our feature learning model, all layers before the last pooling layer of ResNet50 [8] are used as our CNN subnet (see the **red** box in Figure 4). For the FC networks in the hierarchical prediction net (see the **purple** box in Figure 4), we exploit sequential networks with two FC layers followed by ReLU non-linearity layers. The convolutional layers of our CNN subnet are pre-trained on the ImageNet 2012 dataset [22], while the other layers are trained from

Models	$K = 1$	2	3	4	5
NN	34.2	43.6	48.7	52.3	54.0
SGM[7]	31.6	42.5	49.0	53.5	56.8
PPA[30]	33.0	43.1	48.5	52.5	55.4
LSD[2]	33.2	44.7	50.2	53.4	57.6
Ours	**39.0**	**48.9**	**54.9**	**58.7**	**60.5**

Table 1. Comparative results for large-scale FSL on the ImNet dataset, which contains 1,000 source classes. The top-5 classification accuracy (%) over the target class samples is used as the evaluation metric. Our model is shown to significantly outperform the state-of-the-art alternatives. The visualization of these results is shown in Figure 1.

scratch. Stochastic gradient descent (SGD) [13] with momentum is used for model training with a base learning rate of 0.001. For those layers trained from scratch, their learning rate is 10 times the base learning rate. The entire network is trained for 20 epochs on the source class data. The mini-batch size, weight decay, momentum and λ_i (in Eq. 2) are set to 128, 0.0005, 0.9, and 1, respectively.

4.1.2 Comparative Results

The comparative results on the ImNet dataset for large-scale FSL are presented in Figure 1 and Table 1. It can be observed that: (1) Our model outperforms the state-of-the-art large-scale FSL methods, and the improvements are more significant with smaller K. This means that our model is really effective for large-scale FSL. (2) The baseline NN method is competitive and even beats the state-of-the-art when $K = 1$. Note that LSD, SGM, and PPA also freeze the same pretrained ResNet50 for visual feature extraction. This result thus shows that most of the knowledge transfer is done by this initial step and the actual transfer learning methods proposed in LSD, SGM and PPA are not effective with low K values. (3) Without the class hierarchy, our model degenerates to the NN baseline. Ours vs. NN thus serves as the ablation study showing how much the class hierarchy guided feature learning contributes to the superior performance of our model.

4.1.3 Hierarchy Construction with Source Classes

Our model can be easily extended to the class hierarchy built with only source classes. Concretely, we use all source classes as leaves (i.e., class nodes) and then group them into higher-level superclasses, as in Sec. 3.2. Similarly, we recognize target class samples by using the proposed feature learning and label inference models. Figure 5 provides the comparative results obtained by the proposed model using class hierarchies built with only source classes and the hierarchy built with source/target classes. The two hierarchies have the same numbers of superclass layers and number of superclasses in each layer. With these two hierarchies, the proposed model is shown to achieve similar performance.

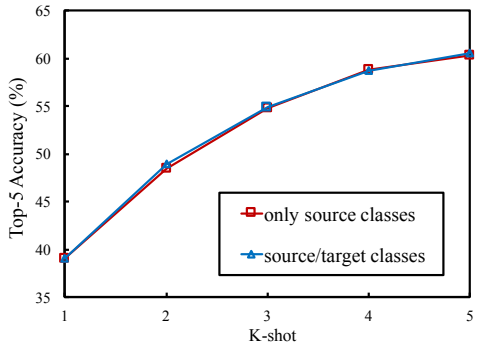

Figure 5. Comparative top-5 accuracy (%) obtained by our model using the hierarchy built with only source classes and the hierarchy built with source/target classes. Our model is shown similar results on both two hierarchies, indicating that our model is effective even without any information from target classes.

No.	n	c_1	c_2	c_3	c_4
1	1	40	–	–	–
2	2	100	10	–	–
3	3	200	40	8	–
4	4	400	160	64	24

Table 2. The details of the four class hierarchies used for selecting the best number of superclass layers on the large-scale ImNet dataset. Notations: n – the total number of superclass layers; c_i – the number of superclasses in the i-th superclass layer (i.e., l_{i+1}).

This suggests that our model can extract transferable features for target class samples *even without any information from target classes*. This is an important result because it shows that the learned feature embedding can generalize to unknown classes. It can be expected that, even without target class names, the semantic relations learned from large text corpus are still encoded into the class hierarchy, making it possible to learn transferable features for target class samples. In other words, the semantic relationship itself is transferable, thus making the learned features transferable.

4.1.4 Hyperparameter Selection for Class Hierarchy Construction

Note that the number of superclass layers and the number of superclasses at each layer are important hyperparameters for class hierarchy construction. Our model follows the recent meta-learning papers [25, 28] to select the hyperparameter values: Firstly, the set of source classes is split into training and validation data/classes; Secondly, the hyperparameters are tuned on the validation data; Finally, we fix the hyperparameters and train our model using all source classes. This hyperparameter strategy determines a 3-superclass-layer class hierarchy, whose three superclass layers respectively have 200, 40, and 8 superclasses for our model. In the following, we provide experimental results to validate the effectiveness of the selected hierarchy structure.

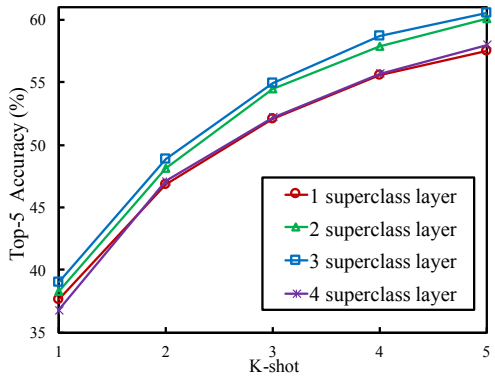

Figure 6. Comparative results obtained by our feature learning model using the hierarchies with different numbers of superclass layers on the large-scale ImNet dataset.

No.	n	c_1	c_2	c_3
1	3	100	20	8
2	3	200	40	8
3	3	400	80	8

Table 3. The details of the three class hierarchies used for selecting the suitable number of superclasses at each superclass layer. The notations are exactly the same as in Table 2.

To validate the effectiveness of the selected number of superclass layers, we construct four class hierarchies with different structures for comparison, and then train our feature learning model with these class hierarchies, as in Sec 3.2. In the same hierarchy, each cluster has similar numbers of nodes on average. The details of the four class hierarchies are given in Table 2. Figure 6 provides the comparative results obtained by our model using the four class hierarchies on the large-scale ImNet dataset. The top-5 classification accuracy (%) on target classes is used as the evaluation metric. It can be observed that the class hierarchy with 3 superclass layers yields the best results.

Given the best number of superclass layers, we then conduct experiments to validate the selected number of superclasses at each superclass layer. Concretely, we first construct two additional class hierarchies with 3 superclass layers for comparison and then train our feature learning model with these two hierarchies. The details of these hierarchies are given in Table 3. The comparative results obtained by our model with these hierarchies on the ImNet dataset are given in Figure 7. It can be seen that the selected number of superclass layers yields the best results. This indicates that the hyperparameter values obtained by cross-validation are indeed optimal.

4.1.5 Further Evaluation

Comparison to Previous Large-Scale FSL Results. Another large-scale FSL setting is adopted by some recent low-shot learning models [7, 30, 2] and thus used here for direct comparison with them. Under this setting, the ImageNet 1K dataset [7] is selected for performance evaluation, with a source/target class split of **389**/611. The visual features are extracted using ResNet50 model [8] trained from scratch with all source samples. We follow the same experimental setup as in [7]: all ImageNet training samples for the source classes, together with K samples per class for the target classes, are available in the training process. We compare our model with six alternatives: NN – nearest neighbor, MN – matching net[28], PN – prototypical net [25], SGM – squared gradient magnitude [7], LSD – large-scale diffusion [2], and PMN – prototype matching net [30]. The top-5 classification accuracy on target classes is used as the evaluation metric. Table 4 provides the comparative results for FSL on the ImageNet1K dataset. It can be seen that our model consistently outperforms the state-of-the-art low-shot learning models [30, 2, 7]. This indicates that the knowledge transfer with class hierarchy induced by our model gives us the edge over existing methods for large-scale FSL. In addition, the NN baseline is shown to be the weakest under this FSL setting. Our explanation is that: only 389 source classes (out of 1000 ImageNet classes) are utilized to train the ResNet50 model, and the obtained feature embedding model is not strong enough to beat the state-of-the-art large-scale FSL models.

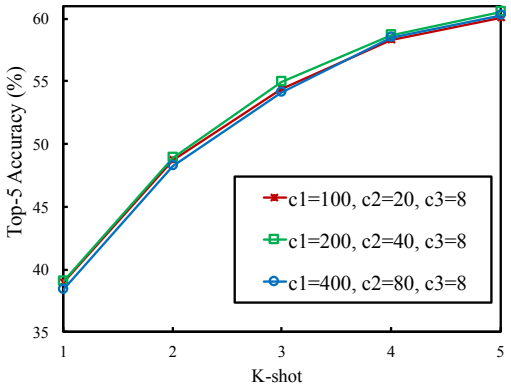

Figure 7. Comparative results obtained by our feature learning model using the hierarchies with different superclass numbers at each superclass layer on the ImNet dataset. Each hierarchy has three superclass layers.

Models	$K=1$	2	5	10	20
NN	49.5	59.9	70.1	75.1	77.6
PN[25]	49.6	64.0	74.4	78.1	80.0
MN[28]	53.3	63.4	72.7	77.4	81.2
SGM[7]	45.1	58.8	72.7	79.1	82.6
LSD[2]	57.7	66.9	73.8	77.6	80.0
PMN[30]	54.7	66.8	77.4	81.4	83.8
Ours	**58.1**	**67.3**	**77.6**	**81.8**	**84.2**

Table 4. Comparative results for large-scale FSL on the ImageNet1K dataset.

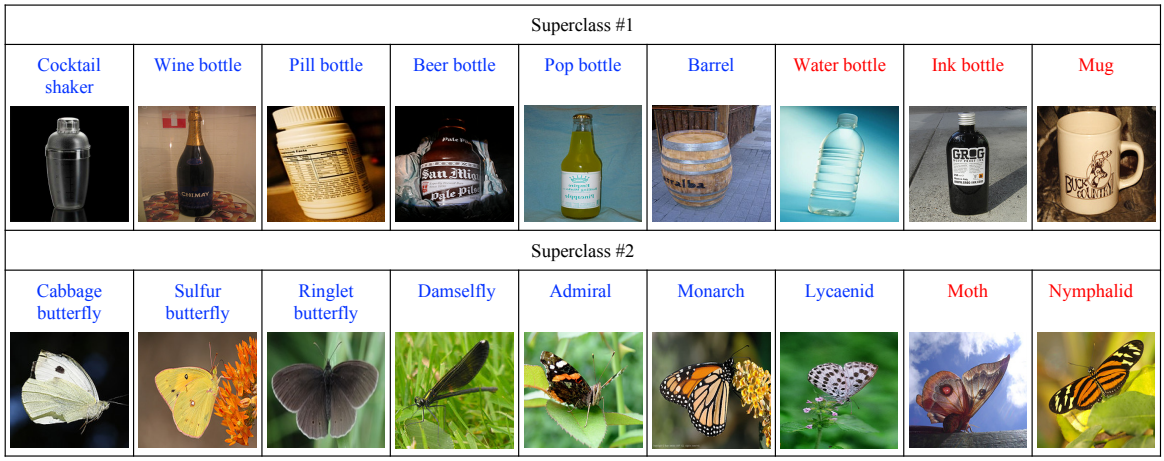

Figure 8. Examples of the superclasses generated by clustering on the ImNet dataset. The source class names are in blue color, while the target class names are in red color. Each class is provided with a visual example. It can be clearly seen that a target class tends to be semantically related to a number of source classes within the same superclass.

Qualitative Results. We provide qualitative results to show why the class hierarchy benefits large-scale FSL. Figure 8 shows two examples of superclasses generated by clustering on the large-scale ImNet dataset. It can be clearly seen that a target class tends to be semantically related to a number of source classes within the same superclass. Since the superclasses are shared across the source and target classes, the encoded class hierarchy itself is transferable, thus making the learned features transferable.

4.2. Large-Scale ZSL

4.2.1 Experimental Setup

In order to assess the suitability of our model for the large-scale ZSL problem, we also run a group of experiments on the large-scale ImNet dataset under the ZSL setting, as in [11]. The only difference from large-scale FSL lies in that no visual samples from target(unseen) classes are provided during the training process, i.e., all samples from unseen classes are used as the test data. The top-5 classification accuracy on all unseen class samples is used as the evaluation metric for large-scale ZSL, as in [11, 5]. We compare our model with the recent and representative ZSL models that have achieved the state-of-the-art results.

4.2.2 Experimental Results

Table 5 presents the comparative ZSL results on the large-scale ImNet dataset. It can be seen that: (1) Our model yields the best results, i.e., it scales well to large-scale ZSL. (2) Our model achieves about 2-5% improvements over the state-of-the-art deep ZSL models [17, 29, 32], showing the effectiveness of our model for solving large-scale ZSL problems. (3) The improvements over the state-of-the-art mapping-learning-based ZSL model [11] demonstrate that the proposed model is more suitable for alleviating the domain gap issue in large-scale ZSL.

Model	Top-5 accuracy (%)
AMP [6]	13.1
SS-Voc [5]	16.8
DeViSE [4]	12.8
ConSE [18]	15.5
VZSL [29]	23.1
CVAE [17]	24.7
DEM [32]	25.7
SAE [11]	27.2
Ours	**27.9**

Table 5. Comparative results for large-scale ZSL on the ImNet dataset. Our model is shown to yield the best results.

5. Conclusion

In this paper, we make an important observation that existing large-scale FSL approaches struggle to beat a simple feature of embedding learning + NN based baseline, indicating their limited scalability and effectiveness. To tackle this problem, we proposed a novel large-scale FSL model by learning transferable visual features with the class hierarchy which encodes the semantic relations between source and target classes. Extensive experiments show that our model achieves state-of-the-art results. Moreover, the proposed model is also shown to achieve promising results in the large-scale ZSL problem.

Acknowledgements This work is supported by National Basic Research Program of China (2015CB352502), NSFC (61573026, 61573363, 61832017), and BJNSF (L172037).

References

[1] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *CVPR*, pages 2927–2936, 2015. 3

[2] Matthijs Douze, Arthur Szlam, Bharath Hariharan, and Hervé Jégou. Low-shot learning with large-scale diffusion. In *CVPR*, pages 7229–7238, 2018. 1, 2, 3, 6, 7

[3] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, pages 1126–1135, 2017. 3

[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*, pages 2121–2129, 2013. 8

[5] Yanwei Fu and Leonid Sigal. Semi-supervised vocabulary-informed learning. In *CVPR*, pages 5337–5346, 2016. 8

[6] Zhenyong Fu, Tao Xiang, Elyor Kodirov, and Shaogang Gong. Zero-shot object recognition by semantic manifold distance. In *CVPR*, pages 2635–2644, 2015. 8

[7] Bharath Hariharan and Ross Girshick. Low-shot visual recognition by shrinking and hallucinating features. In *ICCV*, pages 3037–3046, 2017. 1, 2, 3, 5, 6, 7

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1, 2, 5, 7

[9] Sung Ju Hwang and Leonid Sigal. A unified semantic embedding: Relating taxonomies and attributes. In *NIPS*, pages 271–279, 2014. 3

[10] Michael Kampffmeyer, Yinbo Chen, Xiaodan Liang, Hao Wang, Yujia Zhang, and Eric P. Xing. Rethinking knowledge graph propagation for zero-shot learning. *arXiv preprint arXiv:1805.11724*, 2018. 3

[11] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Semantic autoencoder for zero-shot learning. In *CVPR*, pages 3174–3183, 2017. 1, 2, 3, 5, 8

[12] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 36(3):453–465, 2014. 3

[13] Yann LeCun, Bernhard E. Boser, John S. Denker, Donnie Henderson, Richard E. Howard, and Lawrence D. Hubbard, Wayne E.and Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989. 6

[14] Aoxue Li, Zhiwu Lu, Liwei Wang, Tao Xiang, and Ji-Rong Wen. Zero-shot scene classification for high spatial resolution remote sensing images. *IEEE Trans. Geoscience and Remote Sens.*, 55(7):4157–4167, 2017. 2, 3

[15] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. In *ECCV*, pages 185–201, 2018. 2

[16] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119, 2013. 4

[17] Ashish Mishra, M. Shiva Krishna Reddy, Anurag Mittal, and Hema A. Murthy. A generative model for zero shot learning using conditional variational autoencoders. In *arXiv preprint arXiv:1709.00663*, 2017. 2, 3, 8

[18] 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. 8

[19] Hang Qi, Matthew Brown, and David G. Lowe. Low-shot learning with imprinted weights. In *CVPR*, pages 5822–5830, 2018. 1, 3

[20] Siyuan Qiao, Chenxi Liu, Wei Shen, and Alan L Yuille. Few-shot image recognition by predicting parameters from activations. In *CVPR*, pages 7229–7238, 2018. 1, 2, 3, 5

[21] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *ICLR*, 2017. 1, 3

[22] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 1, 5

[23] Eli Schwartz, Leonid Karlinsky, Joseph Shtok, Sivan Harary, Mattias Marder, Abhishek Kumar, Rogerio Feris, Raja Giryes, and Alex Bronstein. Delta-encoder: an effective sample synthesis method for few-shot object recognition. In *NIPS*, pages 2850–2860, 2018. 3

[24] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[25] Swersky Kevin Snell, Jake and Richard S. Zemel. Prototypical networks for few-shot learning. In *NIPS*, pages 4080–4090, 2017. 3, 6, 7

[26] Richard Socher, Milind Ganjoo, Christopher D. Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. In *NIPS*, pages 935–943, 2013. 3

[27] Sebastian Thrun and Lorien Pratt. *Learning to learn*. Springer Science & Business Media, 2012. 2

[28] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *NIPS*, pages 3630–3638, 2016. 1, 3, 6, 7

[29] Wenlin Wang, Yunchen Pu, Vinay Kumar Verma, Kai Fan, Yizhe Zhang, Changyou Chen, Piyush Rai, and Lawrence Carin. Zero-shot learning via class-conditioned deep generative models. In *AAAI*, 2018. 1, 2, 3, 8

[30] Yu-Xiong Wang, Ross Girshick1, Martial Hebert, and Bharath Hariharan. Low-shot learning from imaginary data. In *CVPR*, pages 7229–7238, 2018. 3, 6, 7

[31] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning - the good, the bad and the ugly. In *CVPR*, pages 4582–4591, 2017. 2, 3

[32] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a deep embedding model for zero-shot learning. In *CVPR*, pages 2021–2030, 2017. 2, 3, 8

[33] An Zhao, Mingyu Ding, Jiechao Guan, Zhiwu Lu, Tao Xiang, and Ji-Rong Wen. Domain-invariant projection learning for zero-shot recognition. In *NIPS*, pages 1027–1038, 2018. 3