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

Ragini Pant[304573]

Department of computer science, Universitätsplatz 1, Universität Hildesheim, 31141 Hildesheim, Germany
pant@uni-hildesheim.de

Abstract. This paper reviews the work done in the field of Large Scale Few-Shot Learning. The Large-Scale FSL model presented in this paper uses Knowledge Transfer with Class Hierarchy. Being a Large-Scale FSL model, it has many instances per class to train on in the source domain but few instances per class in the target domain. The training utilizes pre-trained ResNet-50 as the feature embedding model and a hierarchical prediction net. Classification of the test classes is performed using Nearest Neighbour Search. The approach exploits the transferable visual features learned through the class hierarchy. This approach proposed by the authors significantly beats the baseline approach and other state-of-the-art methods. The feature learning model proposed by the authors can also be extended to Zero-Shot Learning problems and achieve the state of the art results.

Keywords: Few-Shot Learning.· Nearest Neighbour · Zero-Shot Learning · Pre-Trained Model · Image-Classification

1 Introduction

The largest contest in object recognition is the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) held each year, which focuses on the problem of learning 1000 different classes making it a large-scale image recognition problem. A breakthrough occurred when a convolutional neural net won this challenge for the first time with a wide margin, bringing down the state-of-the-art top-5 error rate from 26.2% to 15.3% [18], Since then, these competitions are consistently won by deep convolutional nets, with top-5 error rate in this contest down to 3.6%.[19]

But the Deep Convolution Neural Net based models come with a significant issue of requiring large datasets to train. This means that a lot of image samples have to be collected for each class. And there may be classes that don't have as much data samples as others. So, the focus of the Deep Learning Community has shifted to creating models that do not require as many samples to train and provides an accuracy as good or better than DCNN. This has led to a rise in the field of Meta-Learning which aims at making the model "learn to learn" by

designing networks that can learn new skills and adapt to a new environment with very few training samples.

Few Shot Learning is a sub-topic of Meta-Learning that focuses on learning from a few samples like humans. A FSL task consists of a set of source classes and target classes, with no overlap between them in the label space. The motivation behind the Large-FSL technique is to transfer the knowledge learned from source classes to target classes, along with utilizing the amount of source class samples.

Knowledge transfer from source to target classes to learn transferable visual features is a strong and forgotten baseline for the large-scale FSL setting. It focuses on extracting deep features for target classes and then utilizes it via nearest-neighbor search based classification. Evaluating several latest large-scale FSL models in comparison to this baseline, it was observed that the latest large-scale FSL models ([5],[10],[9],[2],[1]) struggle to beat the baseline, thus displaying their scalability issues. Fig. 1 re-inforces the belief of knowledge transfer, since most of the large-scale FSL models like SGM, LSD, and PPA performed well only because of the transferable features extracted by their deep feature embedding model.

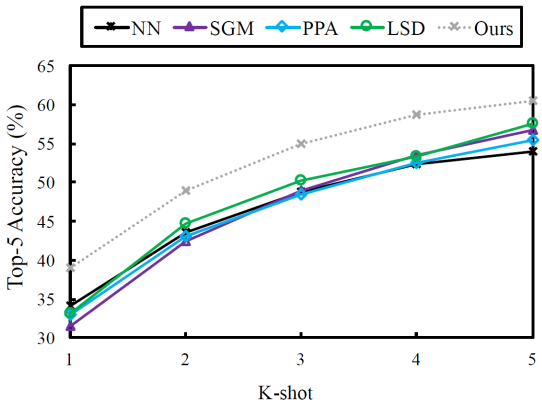

Fig. 1. Comparative Result for the large-scale FSL on the ImNet dataset.Notations: NN –nearest neighbor (NN) search performed in a learned feature space using K samples per target class as the references; SGM – FSL with the squared gradient magnitude (SGM) loss [5]; PPA – parameter prediction from activations (PPA) [7]; LSD – large-scale diffusion (LSD) [2]; Ours – the proposed model.[17]

The novel FSL model proposed by the authors aims at learning as many transferable feature that it can, by learning the semantic relation between source and target classes. The idea is simple and similar to the forgotten baseline approach with an addition of a class hierarchy tree that can encode the semantic relations

between source and target class labels to utilize it as prior knowledge. This will help in producing a better classification accuracy on the test class samples.

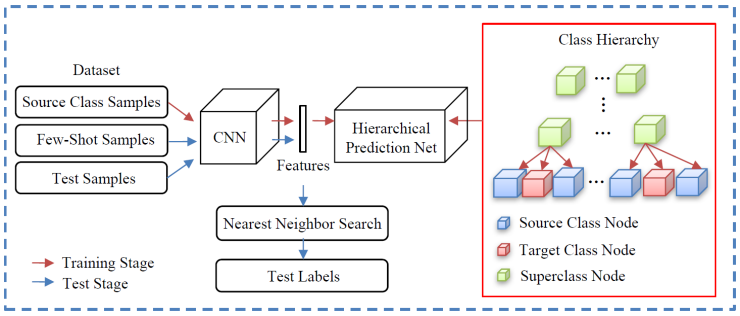

Fig. 2. Overview of the Proposed Model [17]

The authors propose a new feature learning methodology that utilizes Knowledge Transfer with Class Hierarchy. In the hierarchical tree, leaves (nodes) are used to represent source class and target class. These nodes do not overlap with each other, but can share common superclasses. The knowledge generated from the class hierarchy structure is encoded into the network using a prediction net, contributing to better classification result during testing phase. The feature learning model proposed by the authors can also be extended to Zero-Shot Learning Problem.

In this paper, the authors made several **contributions** -

1. Observation of an important baseline that uses deep feature embedding model over source classes for feature extraction and employing this knowledge by performing a nearest neighbor search over target class. The existing large-scale FSL models struggle to beat this baseline revealing their scalability issues.
2. Proposal of a novel approach that uses class hierarchy to learn the semantic relations between source and target classes, and integrate it with the network to generate better classification results. The proposed model outperforms both the Nearest Neighbor (NN) baseline, as well as the state-of-the-art models.
3. Extension of the proposed approach to the Zero-Shot learning problem.

Some information about the authors of this paper :

1. **Aoxue Li** has almost 20 papers from 2015 to 2020.[21]
2. **Tiange Luo** completed his masters from Peking University, and his bachelor from Honors College, Beihang University. His general areas of research include machine perception, generalization, reasoning, and developing computational models with human-like abilities. He has a total of 121 citations[22].
3. **Zhiwu Lu** is a full-time Professor with the School of Information at the Renmin University of China. His educational background includes Masters of Science degree in applied mathematics from Peking University in 2005, and Ph.D. degree in computer science from the City University of Hong Kong in 2011. He has published several papers in many note-worthy conference proceedings and international journals. Some of them include TPAMI, IJCV, TIP, NeurIPS, CVPR, ICCV, and ECCV. He is the winner of IBM SUR Award 2015 and the Best Paper Award at CGI 2014. Along with his team, he stood 2^{nd} in the VID task of ILSVRC 2015. His general areas of interest entail few-shot learning, meta-learning, self-supervised learning, domain adaptation, semi-supervised learning, machine learning theory, monocular 3D object detection, monocular depth estimation, complex video analysis, visual reasoning, face attribute editing, fine-grained object recognition, and image semantic segmentation. He has a total of 1230 citations[23].
4. **Tao Xiang** is currently a Professor of Computer Vision and Machine Learning at University of Surrey, United Kingdom. His educational background includes Bachelor of Science degree in Electrical Engineering from Xi'an Jiaotong University in 1995, Master of Science degree in Electronic Engineering from the Communication University of China in 1998, and Ph.D. Degree in Electrical and Computer Engineering from the National University of Singapore in 2002. He also has a Distinguished Chair at Centre for Vision Speech and Signal Processing (CVSSP). Parallely, he also is a Principal Scientist at Samsung AI Centre, Cambridge, UK where he leads the Multimodal User Interaction Programme. He has also served as an area chair for ICCV'17 and ECCV'20. His general areas of interest comprise of computer vision with a focus on video surveillance, daily activity analysis, and sketch analysis. He also has an interest in large-scale machine learning problems including zero/few-shot learning and domain adaptation. He has a total of 18080 citations [24].

2 Related Work

2.1 Few-Shot Learning

Meta-Learning uses several approaches to tackle the issue of learning from few or no training samples. Some of these approaches are metric-based, optimization-based, predicting parameters of the network, or learning features through hallucination and synthesis of the data sample.

[1] proposes a metric-based approach called Prototypical Networks, that learns the metric space of each class in the latent space via a neural network. Then, the classes are classified by computing the distance of each class to their prototype representation. Another metric-based approach proposed in [2] is called Matching Networks. Here, a new neural-network-architecture utilizes attention mechanisms and memory-efficiency for rapid learning. The optimization-based approach in [3] proposes a meta-learning technique where an LSTM-model can learn an exact optimization algorithm and use it to train another neural-network learner in FSL setting. [4] is a model-agnostic and optimization-based approach, i.e, compatible with any model trained with gradient descent. This approach produces good generalization performance with a small number of gradient steps in the FSL setting. A method proposed in [5] utilizes hallucination and shrinking of samples on the ImageNet1k dataset. It also introduces a new benchmark that focuses on feature representation and low-shot learning on base and novel classes. Apart from hallucination, new samples can be synthesized using an autoencoder and then further trained using these new data samples as proposed in [6]. Meta-learning techniques also utilize information gained from parameters and activations of a network.[7] proposes a method that can adapt the weights using activations of a pre-trained neural network to classify test classes. Parameter weight imprinting is proposed in [8], where the weights for a novel category is set using the latent layer activations for its target class. This method indicates better results than nearest-neighbor instance embedding.

Most of these techniques are trained on Mini-ImageNet, CIFAR, or Omniglot dataset and don't utilize the number of data samples available for source class. [5] and [7] focus on Large-Scale FSL, thus utilizing the source class data samples. [7] experimented on both Full-ImageNet and Mini-ImageNet, and achieved state-of-the-art classification accuracy on novel categories by a significant margin while maintaining comparable performance on the large-scale classes.[10] proposes large-scale low-shot learning combined with a hallucinator to produce more data samples, and provide substantial gains in the classification accuracy. Another technique that combines the diffusion method in a semi-supervised learning setting over large-scale is proposed in [9].

2.2 Zero-Shot Learning

Zero-Shot learning assumes no visual data sample for the target class. Rather, only textual attributes of the target class are provided to help in the knowledge transfer of the features and help with the classification.[11] and [12] define an approach that can directly learn to map the visual image space to a semantic space, with only source class data leading rise to domain gap issue between seen and unseen classes. Though, this is possible because these approaches do not exploit the semantic relation between the source and target classes like [17] and learn the transferable features.

2.3 Knowledge Transfer with Class Hierarchy

In FSL and ZSL, little focus on the importance of class hierarchy in knowledge transfer is put. [15] and [16] are the only exceptions that utilize class-hierarchy. [16] uses class hierarchy to define a semantic space for ZSL, whereas [15] employs the class hierarchy to learn the relation between categories, supercategories, and attributes. [15] manually-defines the class-hierarchy, thus adding additional cost to the experiment. However, [16] expresses that manually-defining the class hierarchy is an expensive task, so the hierarchical embedding is derived from WordNet text corpus.

3 Summary

3.1 Problem Definition

The problem is defined by the authors as follows: [17]

1. S_{source} denotes the set of source classes.
2. S_{target} denotes the set of target classes.
3. Both of these sets are disjoint, i.e $S_{source} \cap S_{target} = \phi$.
4. In the experiment, the authors assume that they are given a large-scale sample set D_{source} from source classes S_{source}, and a few-shot sample set D_{target} from target classes S_{target}.
5. Here, K $\leq$ 5 i.e number of training samples per class in D_{target}.

3.2 Transferable Visual Feature Learning

The authors proposed a novel transferable feature learning model for large-scale FSL.

1. To encode the semantic relations between source and target classes, a tree structures class hierarchy is constructed.
2. Integration of the knowledge of the class-hierarchy into network is done using a hierarchical prediction net.
3. Training the model on D_{source} and the semantic data that represents the relationship between source and target data labels, i.e the text corpus.

3.2.1 Tree structured class hierarchy

The hierarchical tree structure proposed by the authors, utilizes the source and target classes as the leaves of the tree (classes), thus forming the bottom layer. The tree-like structure can be achieved by training a skip-gram model on a corpus of 4.6M Wikipedia documents and then representing each source/target class label using a word vector extracted from this skip-gram model. By starting from the bottom, similar word vectors can be clustered together using k-means clustering, forming the bottom layer of the class-hierarchical structure. Similarly,

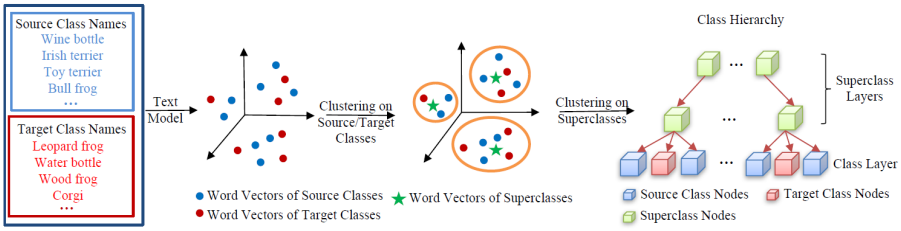

Fig. 3. Illustration of tree-structured class hierarchy [17]

clustering over the nodes of the bottom layer, an upper-class node is created. These upper-class nodes are also called a parent node or superclass node, which together form a superclass layer. This approach generates a tree structure class hierarchy with one normal class layer (bottom layer) and n superclass layer.

3.2.2 Hierarchical Prediction Net

The hierarchical prediction net predicts only the superclass labels since both source and target class labels are utilized as the leave nodes. Fig.4 describes the operation of the proposed model. Features are extracted using a pre-trained ResNet50. These generated features are used by the prediction net, First, labels at different class/ superclass layer are predicted. This indicates the transferable visual feature learning. (as denoted in the yellow box in Fig.4) In the second step, the authors integrate the hierarchical structure of class/ superclass into superclass label prediction.(as shown in the **green box** in Fig.4) To infer the labels of each superclass in each layer, the prediction results of the same and lower class/superclass layers are combined.

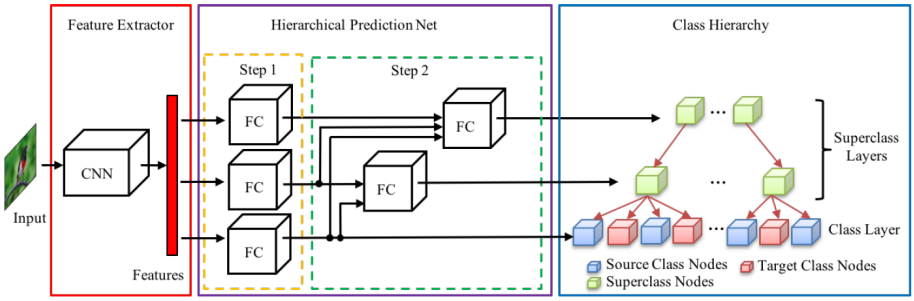

Fig. 4. Overview of the proposed feature learning model. Notation: 'FC'– fully-connected network.[17]

In Figure 4 -

1. The yellow dashed box acts as the first-class/superclass label prediction step. Here, the authors have added *n + 1* unshared fully-connected (FC) networks with softmax layers on top of the CNN model that helps to predict the probability distribution for each class.

2. The green dashed box acts as the second superclass label prediction step, where the authors have integrated the hierarchical structure among the class /superclass layers by using n unshared FC networks, each inferring the labels of the corresponding superclass layer.

3. For the FC network corresponding to the lowest superclass layer, inference of labels of the corresponding superclass layer is performed by concatenating the outputs of the bottom two layers (Step 1 of Prediction Net) and feed it as input to the FC network in Step 2 of the Prediction Net. The output of this FC network is the final prediction result for the lowest superclass layer. (as shown in Eq. 1)

$$\hat{p}_{l2} = \mathcal{F}_{l_2}^2(p_{l1} \oplus p_{l2}) \tag{1}$$

   here,
   - p_{l1} and p_{l2} denotes the output of the bottom FC network and the second FC network respectively.
   - $\oplus(\cdot)$ denotes a concatenation operator.
   - $\mathcal{F}_{l_2}^2(\cdot)$ denotes a forward step of the FC network corresponding to the layer l_2 in Step 2 of the Prediction Net.
   - The output $\hat{p}_{l2}$ denotes the final predicted distribution over all possible superclass labels at the second layer of the class-hierarchy.

4. Inference of superclass labels from third layer is performed by concatenating the outputs of the current layer and the lower layers (Step 1 of the Prediction Net). This is fed as input to the FC network for the corresponding layer in Step 2 of the Prediction Net. The authors define the loss function for an image x by merging the hierarchical inference steps with the class label prediction step.

$$p_{li} = \mathcal{F}_{l_i}^1(\mathcal{G}(x)) \ , \ i = 1,....,n+1$$

$$\hat{p}_{li} = \mathcal{F}_{l_i}^2(\oplus_{j=1}^{i} p_{lj}), i = 2, ..., n+1 \tag{2}$$

$$\mathcal{L}(x, \mathcal{Y}; \ominus) = \mathcal{L}_{cls}(y_{l1}, p_{l1}) + \sum_{i=2}^{n+1} \lambda_i \mathcal{L}_{cls}(y_{li}, \hat{p}_{li})$$

   here
   - G denotes the forward step of the CNN for feature extraction.
   - $\mathcal{F}_{l_i}^1$ and $\mathcal{F}_{l_i}^2$ denote a forward step of the FC network corresponding to layer l_i in Step 1 and Step 2 of the prediction net respectively.

- p_{li} denotes the predicted distribution over all possible classes/superclasses in layer l_i in Step 1 of the prediction net.
- $\hat{p}_{li}$ denotes the final predicted distribution over all possible superclasses in layer l_i.
- $\oplus$ denotes the concatenation operator.
- Y = y_{li},i = 1,....,n+1 collects the true class/superclass labels of the image x, where y_{li} denotes the label corresponding to layer l_i
- $\ominus$ denotes the parameters of the full network.
- $\mathcal{L}_{cls}$ denotes the cross entropy loss for classification.
- λ_i denotes the regularization hyperparameter.

3.3 Label Inference

After the 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}) using a nearest-neighbor search method that utilizes the visual features obtained from the deep-feature embedding model.

1. To infer the labels of test samples from D_{test}, compute the cosine distance to each class reference, and predict the class label as the one with the smallest distance.
2. To infer the labels of target samples from D_{target}, compute the average of the visual features of its few-shot samples as its reference.

3.4 Extension to Large Scale ZSL

The authors discovered that the proposed feature learning model, which was originally designed for large-scale FSL, can be easily extended to large-scale ZSL. Here, the training data comprises of S_{source} but no visual samples from S_{target}, just the labels. Similar to large-scale FSL, a tree-structures class hierarchy is constructed using the word vectors of all the S_{source} and S_{target} class names. With the obtained class hierarchy, the deep feature learning model is trained over the whole training set, i.e S_{source}. Using state-of-the-art-mapping-learning [12], the visual features extracted from the deep-feature embedding model can be used to infer the labels of test images. Experiments were performed on the ImNet dataset, and this proposed model creates a new stare-of-the-art Large-scale ZSL model.

4 Experimental Setup and Analysis

4.1 Large-Scale FSL

Dataset : For the experiments and performance evaluation the authors derived a new benchmark dataset from ILSVRC2012/2010 (ImNet).This dataset is organized as below :

1. S_{source} : A training set of many labeled source class samples, comprising of 1000 classes of ILSVRC2012.

2. S_{target} : A few-shot set of few labeled target class samples, from the 360 classes of ILSVRC 2010 (not included in ILSVRC2012)
3. Target classes not used, as the test-set.

Model Settings : For the deep feature embedding model, the authors used ResNet50 pre-trained on ILSVRC2012 1K classes. Then, for the classification over the target samples, the authors used a simple nearest-neighbor model.

Other Benchmark Models : The other models that the authors used for comparison were :

1. **NN -** This is the nearest neighbor search based model. It is used to generate a strong baseline in feature space using K samples per target class, with ResNet50 as the pre-trained feature-embedding model.
2. **SGM -** It is a meta-learning technique based on shrinking and hallucinating the features of a model. It also introduces a new benchmark that focuses on feature representation and low-shot learning on base and novel classes.
3. **PPA -** A few-shot learning model that uses parameter prediction with activations model for better performance on a few training samples.
4. **LSD -** This is also a few-shot learning model which utilizes large scale diffusion method in a semi-supervised learning setting.

Evaluation Metric : For evaluation with other models, the top-5 accuracy on the overall test images is computed for each sample.

Network Architecture and Training Details : The class hierarchy model is a 3-superclass-layer, consisting of 200, 40, and 8 superclasses, respectively created by k-means clustering over the word vectors of source/target class names. In the feature learning model -

1. The CNN subnet is a pre-trained ResNet50, except the last pooling layer. (**red box** in Fig.4)
2. In the hierarchical prediction net, ReLU activation is used after the two FC layers. (**purple box** in Fig.4)
3. ResNet50 is pre-trained on S_{source}, while the other layers are trained from scratch.
4. The optimization technique is Stochastic gradient descent (SGD) with momentum.
5. The base learning rate is 0.01, and the learning rate for layers trained from scratch is 0.001.
6. Other hyperparameters like mini-batch size, weight decay, momentum, and i (in Eq. 2) are set to 128, 0.0005, 0.9, and 1, respectively.
7. The entire model is trained on S_{source} for 20 epochs.

4.1.2 Comparative Results

Figure 1 and Figure 5 show the comparative results for the large-scale FSL models on the ImNet dataset. The results indicate that -

1. The proposed model outperforms the benchmark large-scale FSL models, with the best performance achieved with a smaller value of K.
2. The NN baseline beats the state-of-the-art models when K = 1. Since, LSD, SGM, and PPA also use the same pre-trained ResNet50 for visual feature extraction, it can be concluded that most of the knowledge transfer occurs at the feature extraction step and the actual transfer learning methods proposed in LSD, SGM and PPA are not effective with low K values.
3. Without the proposed class hierarchy, there is no difference between the NN baseline and the proposed model by the authors, indicating the contribution of class hierarchy guided feature learning makes the model perform superior.

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**

Fig. 5. Comparative results for large-scale FSL on the ImNet dataset. The visualization of these results is shown in Figure 1.[17]

4.1.3 Hierarchy Construction with Source classes

The class hierarchy can be built with both source/target classes or just source classes as well. Fig. 6 show the result of the experimentation with both class-hierarchy constructions, built with the same number of superclass layers and the same number of superclasses in each layer. The conclusion is that the model shows similar result on both two hierarchies, indicating that the model is effective and it derives enough knowledge from the source class and can learn enough feature embeddings to generalize well to unknown data. The authors suggest that this might be possible because, while constructing the class-hierarchy structure the semantic relations between source and target classes were encoded.

4.1.4 Hyperparameter Selection for Class Hierarchy Construction

To select the optimal value for the hyperparameters like the number of superclass layers and the number of superclasses in each layer, cross-validation technique is used. The result obtained is 3 for the number of superclass layers, and each superclass layer to have 200, 40, and 8 superclasses, respectively.

The authors validated the effectiveness of the selected number of superclass layers by constructing a 4-superclass-layer class hierarchy with different structures and then further training these structures on the feature learning model. After evaluating the best number of superclass layers, the authors also conducted ex-

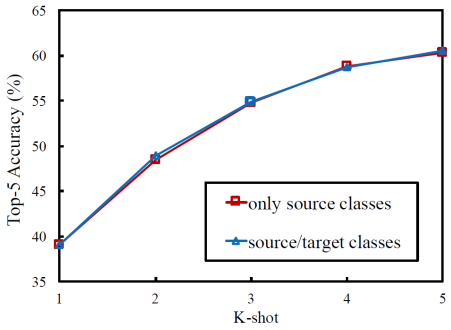

Fig. 6. Comparative top-5 accuracy (%) obtained by the model using the hierarchy built with only source classes and the hierarchy built with source/target classes.[17]

periments to validate the number of superclasses at each superclass layer. The details of these structures are provided in Fig.7 and Fig.8. The performance was evaluated on these different structures as shown in Fig.9 and Fig.10.This indicates that the hyperparameters obtained through cross-validation were optimal.

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

Fig. 7. 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}).[17]

4.1.5 Further Evaluation

Comparing with Previous Large-Scale FSL Results : Figure 11 exhibit the comparison between some recent low-shot learning models with the proposed model by the authors is performed. The experimental setup is same as in the above case, except the source/target class split is different. There are 389 classes in source class and 611 classes in the target class. The performance of the proposed model is not optimal in this setting, since the number of source classes are less, hence the obtained deep feature embedding is not strong enough.

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

Fig. 8. The details of the three class hierarchies used for selecting the suitable number of superclasses at each superclass layer.[17]

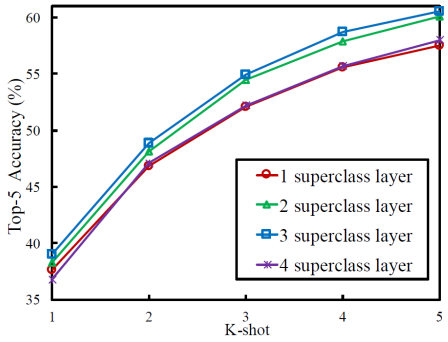

Fig. 9. Comparative results obtained by the feature learning model using the hierarchies with different numbers of superclass layers on the large-scale ImNet dataset.[17]

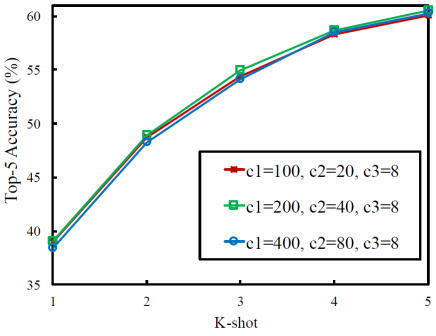

Fig. 10. Comparative results obtained by the feature learning model using the hierarchies with different superclass numbers at each superclass layer on the ImNet dataset. Each hierarchy has three superclass layers.[17]

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**

Fig. 11. Comparative results for large-scale FSL on the ImageNet1K dataset.Notation: 'NN'– nearest neighbor[17], 'MN'– matching net [2], 'PN'– prototypical net [1], 'SGM'– squared gradient magnitude [5], 'LSD'– large scale diffusion [9], and 'PMN'– prototype matching net [10]

Qualitative Results Figure 12 depicts examples of superclasses generated through k-means clustering on the text corpus. It indicates that within the same superclass, the semantic relation between target and source classes is very strong.

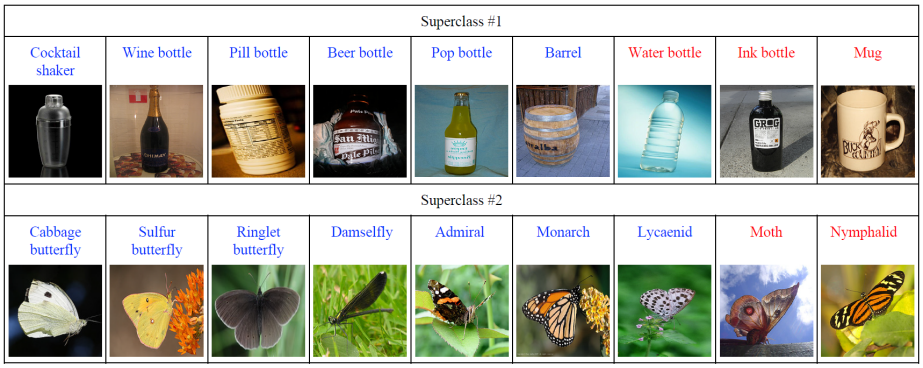

Fig. 12. Examples of the superclasses generated by clustering on the ImNet dataset. The source class names are in blue color, while the target class names are in red color. Each class is provided with a visual example.[17]

4.2 Large Scale ZSL

4.2.1 Experimental Setup

Using the experimental setup for Large-Scale ZSL from [13] the authors run a group of experiments on the large-scale ImNet dataset. In the ZSL setting, the target class samples are utilized as test class samples.

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**

Fig. 13. Comparative results for large-scale ZSL on the ImNet dataset.[17]

4.2.2 Experimental Results

Figure 13 shows the result of the proposed model with other state-of-the-art ZSL models on the ImNet dataset, establishing that the proposed model outperforms all other state-of-the-art Large-scale ZSL models. The results signify that -

1. The proposed model yields the best result in the ZSL setting as well.
2. The model achieves about 2-5% improvements over the state-of-the-art deep ZSL models, indicating the effectiveness of class hierarchy in learning transferable visual features even though no visual samples of target class is provided. Thus, alleviating the domain gap issue.

5 Discussion

The authors' motivation behind large-scale FSL with class hierarchy is to extract similar visual features between source and target class and use them in a hierarchical tree structure for better knowledge transfer. Learning all the knowledge from the class hierarchy during the training will help the model predict test samples with high accuracy. Another motivation is to create an inexpensive model, so the authors create a class-hierarchical structure without human annotation.

The authors did not explicitly formulate the Research Questions behind the paper, but it can be implied that the main questions were -

1. How to create a Large-Scale FSL model that can eradicate the scalability issues faced in the current state-of-the-art approaches?
2. How to increase the knowledge transfer from source to target class, that can help in learning the visual feature representation of the classes?
3. Is it possible to extend the current motivation to Zero-Shot-Learning Setting, where no visual samples for the target class are provided?
4. Can the hierarchical tree generated from class labels, act as a good prior for feature extraction?

5. How to encode the class hierarchy in the classification procedure?

The Related Work section mentioned in the paper is comprehensive, since it mentions all the different meta-learning approaches like metric-based, optimization-based, using parameters and activations of the network and feature-learning through hallucination and synthesis of the data sample. But, the authors do not explain the principle ideas behind these papers. Also, the benchmark Large-Scale FSL techniques were not extensively explained. In the Related Work section for Zero-Shot Learning, the authors again mention the state-of-the-art models, but the core idea behind them is not described. The authors use the class-hierarchy idea from [15] and [16]. The purpose of class hierarchy in each of these papers in explicitly stated. In [17], the authors also mention that [16] uses a human-annotated class hierarchy; however, in my understanding, the class hierarchical embedding was derived using the WordNet corpus.

The methodology section of the paper for Large-Scale FSL is extensive. The Large-Scale FSL problem is formally defined. The authors also provide a detailed description of the construction of a tree-structured class hierarchy, as well as the prediction net utilized by the network for the prediction of correct classes. Other parts of the model, like the pre-trained model used activation functions and other training procedure were defined as well. Also, all the formulas described are logically correct. Though the superclass loss uses L1 regularization parameter, and the authors have not explained the use-case of it. The methodology section for Large-Scale ZSL is not very detailed and lacks the explanation for the training procedure of the large-scale ZSL model. The label inference in Large-Scale ZSL setting is done using a state-of-the-art mapping which is not described as well.

The proposed model has several strengths that are worth mentioning, like -

1. Using a hierarchical tree generated from class labels acts as a good prior for classification.
2. A similar approach could be easily extended to the Zero-Shot Learning setting.
3. This is an inexpensive approach since the hierarchy tree did not have to be human-annotated.
4. The results achieved through this method were able to remove the scalability issue that was faced by other benchmark models.
5. The proposed model was able to achieve state-of-the-art results with a huge margin.

Though the results were really good. The authors could have tried some other approaches to increase the performance even further, like -

1. The authors could have tried constructing the hierarchical tree using different word embedding methods, and reported the results.

2. The authors used k-means clustering for the word vectors to create nodes in the hierarchical tree. They could have tried kernel-based clustering methods and reported the results.
3. The authors could have tried to report the change in loss, without regularization and with L1 and L2 regularisation.

In the Experimental Setup for the proposed Large-Scale FSL model, the novel dataset used by the authors is described. Also, the model settings, network architecture, and training details are mentioned. The authors used top-5 accuracy on overall test images as the evaluation, which is a good evaluation metric considering the large size of the dataset. However, the authors did not provide a thorough description of the Large-Scale ZSL experimental settings.

The results for all the experiments on the hyperparameter - selection, hyperparameter - tuning and hierarchical construction with source and target classes are neatly presented with tables and graphs. The comparative result for the proposed Large-Scale FSL model with the current state-of-the-art models on the ImNet dataset is represented well, using charts and graphs. Similarly, the comparative result for large-scale ZSL with the current benchmark models on the ImNet dataset is represented, though using just a table. The authors should have included a graph for it too.

In the Conclusion section of the paper, the authors indicated the poor-performance of the current benchmark models in comparison to the baseline model. They also established that their proposed model alleviates all the disadvantages incurred by the benchmark models, and even performs better than both baseline model and the benchmark models. The authors were able to answer all the implicit research questions with a positive outcome. With the result obtained on both Large-Scale FSL and Large-Scale ZSL, it can be concluded that the learning the semantic relation between source and target class in a class hierarchy, helps in learning good transferable visual features and leads to more knowledge transfer.

The paper is formally correct, presented in a precise and clear format with proper grammar and spelling. All the graphs, figures, tables, and citations are accurate. The paper was introduced in 2019 in the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition and has been cited 19 times[20]. It is the top conference for the Computer Vision and Pattern Recognition category with the h5-index of 299.

6 Conclusion

In the paper, the authors have identified the limited scalability and effectiveness issue of the current-state-of-the-art Large-Scale FSL models. They have also highlighted the poor performance of the benchmark models against the baseline model, i.e feature embedding learning with NN approach. To alleviate the problems of existing state-of-the-art-models that do not utilize the knowledge

of transferable visual features, the authors proposed a novel FSL model that focuses on learning these feature embeddings with class hierarchy. The class hierarchy encodes semantic relations between source and target classes and achieves promising results in both large-scale FSL and large-scale ZSL problems. This highlights the importance of learning transferable visual features to identify the target class in large-scale FSL and large-scale ZSL problem settings.

Bibliography

[1] Snell, Jake, Kevin Swersky, and Richard Zemel. "Prototypical networks for few-shot learning." Advances in neural information processing systems. 2017.

[2] Vinyals, Oriol, et al. "Matching networks for one shot learning." Advances in neural information processing systems. 2016.

[3] Ravi, Sachin, and Hugo Larochelle. "Optimization as a model for few-shot learning." (2016).

[4] Finn, Chelsea, Pieter Abbeel, and Sergey Levine. "Model-agnostic meta-learning for fast adaptation of deep networks." arXiv preprint arXiv:1703.03400 (2017).

[5] Hariharan, Bharath, and Ross Girshick. "Low-shot visual recognition by shrinking and hallucinating features." Proceedings of the IEEE International Conference on Computer Vision. 2017.

[6] Schwartz, Eli, et al. "Delta-encoder: an effective sample synthesis method for few-shot object recognition." Advances in Neural Information Processing Systems. 2018.

[7] Qiao, Siyuan, et al. "Few-shot image recognition by predicting parameters from activations." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

[8] Qi, Hang, Matthew Brown, and David G. Lowe. "Low-shot learning with imprinted weights." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[9] Douze, Matthijs, et al. "Low-shot learning with large-scale diffusion." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

[10] Wang, Yu-Xiong, et al. "Low-shot learning from imaginary data." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[11] Xian, Yongqin, Bernt Schiele, and Zeynep Akata. "Zero-shot learning-the good, the bad and the ugly." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.

[12] Lampert, Christoph H., Hannes Nickisch, and Stefan Harmeling. "Attribute-based classification for zero-shot visual object categorization." IEEE transactions on pattern analysis and machine intelligence 36.3 (2013): 453-465.

[13] Kodirov, Elyor, Tao Xiang, and Shaogang Gong. "Semantic autoencoder for zero-shot learning." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.

[14] Zhao, An, et al. "Domain-invariant projection learning for zero-shot recognition." Advances in Neural Information Processing Systems. 2018.

[15] Hwang, Sung Ju, and Leonid Sigal. "A unified semantic embedding: Relating taxonomies and attributes." Advances in Neural Information Processing Systems. 2014.

[16] Akata, Zeynep, et al. "Evaluation of output embeddings for fine-grained image classification." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[17] Li, Aoxue, et al. "Large-scale few-shot learning: Knowledge transfer with class hierarchy." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.

[18] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012.

[19] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. Deep Learning. The MIT Press.

[20] Citation counts and Author Information is taken from https://scholar.google.com/. Accessed on 15-August-2020.

[21] Aoxue Li : https://dblp.uni-trier.de/pers/l/Li:Aoxue.html (Accessed on 15-August-2020)

[22] Tiange Luo : https://tiangeluo.github.io/ (Accessed on 15-August-2020)

[23] Zhiwu Lu : https://sites.google.com/site/zhiwulu/ (Accessed on 15-August-2020)

[24] Tao Xiang : http://personal.ee.surrey.ac.uk/Personal/T.Xiang/index.html (Accessed on 15-August-2020)