Fusing Knowledge and Experience with Graph Convolutional Network for Cross-task Learning in Visual Cognitive Development

Xinyue Zhang[1(✉)], Xu Yang[1], Zhiyong Liu[1], Lu Zhang[1], Dongchun Ren[2], and Mingyu Fan[2]

[1] State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, People's Republic of China
{Zhangxinyue2020,xu.yang}@ia.ac.cn

[2] Meituan-Dianping Group, Beijing 100190, People's Republic of China
rendongchun@meituan.com

Abstract. Visual cognitive ability is important for intelligent robots in unstructured and dynamic environments. The high reliance on large amounts of data prevents prior methods to handle this task. Therefore, we propose a model called knowledge-experience fusion graph (KEFG) network for novel inference. It exploits information from both knowledge and experience. With the employment of graph convolutional network (GCN), KEFG generates the predictive classifiers of the novel classes with few labeled samples. Experiments show that KEFG can decrease the training time by the fusion of the source information and also increase the classification accuracy in cross-task learning.

Keywords: GCN · Few-shot learning · Cognitive development · Transfer learning · Image recognition · Cross-task learning

1 Introduction

Visual cognitive development is vital for intelligent robots in unstructured and dynamic environments. However, limited by the number of available labeled samples, an intelligent robot faces lots of novel categories in realistic environments. An ideal robot can transfer the knowledge and experience from the base classes to novel ones independently. The ability of predictive knowledge transferring efficiently cuts down the training cost and extends the range of cognition. Since

This work is supported partly by National Key R&D Program of China (grants 2017YFB1300202 and 2016YFC0300801), partly by National Natural Science Foundation (NSFC) of China (grants 61973301, 61972020, 61633009, and U1613213), partly by Beijing Science and Technology Plan Project (grant Z181100008918018), and partly by Meituan Open R&D Fund.

F. Sun et al. (Eds.): ICCSIP 2020, CCIS 1397, pp. 114–120, 2021.
https://doi.org/10.1007/978-981-16-2336-3_10

the robot only obtains a few labeled samples of the novel categories other than some supplementary information, few-shot learning is often used.

Previously, the basic idea of few-shot learning is to exploit the knowledge from the base classes to support the novel ones. Most existing methods can be divided into two groups: metric-learning based method [1,8,9] and meta-learning method [10–12]. However, these methods mainly focus on inter-class dissimilarity. The relationship among classes is also important for knowledge transfering. To improve the efficiency of the sample utilization, some researchers exploit the knowledge graph to make up a relation map among categories [2,3,5]. Thus the structural information is also taken into account and the propagation becomes more reasonable. However, pure knowledge information shows a deviation between semantic space and visual space. This one-sidedness of pure knowledge graph leads to an unsatisfactory accuracy. Besides, the amount of knowledge graph to support the novel classes is huge, due to the sparseness of the information. Thus a more sophisticated inference mechanism is in need.

Humans can rapidly adapt to an unfamiliar environment. This is mainly based on two parts of information: *common sense* and *visual experience*. With the common sense, they know the descriptions of the novel class and its relationship with the base ones. With the experience of learning the base classes, they quickly grasp the method to classify the novel ones. These two parts of information enable humans to develop their visual cognitive ability accurately and quickly.

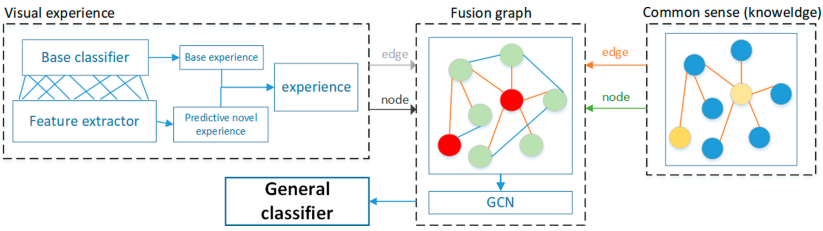

Fig. 1. We jointly explore the source information from visual experience and common sense to predict the general classifier of novel categories.

Motivated by this, we propose a model called knowledge-experience fusion graph network (KEFG) for few-shot learning. The goal of KEFG is to jointly explore the common sense (knowledge) and visual experience to accomplish the cognitive self-development of robots. For convenience and according to the daily usage, below the common sense is directly abbreviated by *knowledge*, while the visual experience is denoted by *experience*. Specifically, KEFG obtains experience from the original trained recognition model based on Convolutional Neural Network (CNN). It recalls the visual representation of the base classes and generates the predictive classifiers of the novel ones. KEFG further explores the prestored knowledge graph from WordNet [7] and builds a task-specific subgraph for efficiency. With the employment of GCN [4], novel classes generate its own classifier

following the mechanism of related base classes on the fusion graph. To evaluate the effectiveness of KEFG, cross-task experiments are conducted to transfer the cognition ability from ImageNet 2012 to two typical datasets, fine-grained medium size dataset Caltech-UCSD Birds 200 (CUB) [13] and coarse-grained small size dataset miniImageNet [8]. The results show satisfactory performance.

The main contributions are as follows: (1) The knowledge graph builds a developmental map suitable for cognitive development. KEFG conducts a developmental framework to transfer the base information to specific tasks. (2) KEFG jointly explores information from the visual space and word space. It cuts down the number of nodes to support the inference and decreases the deviation. (3) The experiments show that KEFG conducts well not only on the coarse-grained small size dataset but also on the fine-grained medium size dataset.

2 Methodology

The set of all categories contains training set C_{train}, support set $C_{support}$, and testing set C_{test}. C_{train} has sufficient labeled images. $C_{support}$ and C_{test} are from the same categories called *novel classes*, while the training categories called *base classes*. If the support set contains K labeled samples for each of the N classes, we call this problem $N - way\ K - shot$ few-shot problem. KEFG is built on an undirected knowledge graph, denoted as $G = (V, E)$. V is a node set of all classes. Each node represents a class. $E = \{e_{i,j})\}$ is an edge set. The classification weights are defined as $w = \{w_i\}_{i=1}^{N}$ where N is the number of total categories.

2.1 Information Injected Module

KEFG employs the knowledge graph from WordNet. Better than taking the whole graph, KEFG adds the novel classes to the constant graph of the base classes in ImageNet 2012. If there are N classes, KEFG takes a subgraph with N nodes. To transfer the description into vectors, we use the GloVe text model and get S input features per class. The feature matrix of knowledge is $V_K \in R^{N \times S}$. KEFG only uses the hyponymy as the principle of the edge construction. The edge matrix from the knowledge space refers to $E_K \in R^{N \times N}$.

KEFG learns from the experience of the original model which is denoted as $C(F(\cdot|\theta)|w)$. It consists of feature extractor $F(\cdot|\theta)$ and category classifier $C(\cdot|w)$. θ and w indicate the classification parameters of the model. Feature extractor $F(x|\theta)$ takes an image as input and figures out the feature vector of it as z_i. The parameter w^{train} refers to the classification weights of different classes in training set. The final classification score is computed as $s = \{z^T w\}$

The feature extractor part $F(\cdot|\theta)$ can compute feature representations of the $C_{support}$. According to the rule of the template matching, the feature representation of the novel class can well represents its classification weights. Thus the initial weights can be represented as the average of the features.

$$v_i^E = \begin{cases} w_i^{train}, & x_i \in C_{train} \\ \frac{1}{P} \sum_{k=1}^{P} F(x_{i,p}|\theta), & x_i \in C_{test} \end{cases} \tag{1}$$

Where v_i^E refers to the visual feature of the ith class. $x_{i,p}$ refers to the pth image in the ith class. P is the total number of the images in the ith class.

Motivated by the denoising autoencoder network, KEFG injects the word embedding to the initial classification weights to generate a more general classifiers. The features of the ith class in the fusion graph is represented as follows

$$v_i = \alpha \frac{v_i^K}{\|v_i^K\|_2} + \beta \frac{v_i^E}{\|vi^E\|_2} \tag{2}$$

where α and β refers to the proportion of each source of information.

Except the relationship of hyponymy, KEFG also introduces cosine similarity to the graph. The edges are denoted as follows

$$e_{i,j} = \begin{cases} 1, & Simi(x_i, x_j) > S \text{ or } Hypo(x_i, x_j) \\ 0, & otherwise \end{cases} \tag{3}$$

$Simi(x_i, x_j)$ refers to the cosine similarity. S represents the similarity boundary to judge whether there is a relationship between two classes and it is a hyperparameter. $Hypo(x_i, x_j)$ refers to the mechanism to judge whether there if the relationship of hyponymy between ith class and jth class.

2.2 Information Transfer Module

With the framework of the GCN, KEFG propagates information among nodes by exploring the classes relationship. The mechanism of GCN is described as

$$H^{(l+1)} = ReLu(\hat{D}^{-\frac{1}{2}}\hat{E}\hat{D}^{-\frac{1}{2}}H^{(l)}U^{(l)}) \tag{4}$$

where $H^{(l)}$ denotes the output of the lth layer. $\hat{E} = E + I$, where $E \in R^{N\times N}$ is the symmetric adjacency matrix and $I \in R^{N\times N}$ represents identity matrix. $D_{ii} = \sum_j E_{ij}$. U^l is the weight matrix of the lth layer.

The fusion graph is trained to minimize the loss between the predicted classification weights and the ground-truth weights.

$$L = \frac{1}{M}\sum_{i=1}^{M}(w_i - w_i^{train})^2 \tag{5}$$

where w refers to the output of base classes on GCN. w^{train} denotes the ground truth obtained from the category classifier. M is the number of the base classes.

KEFG further applies the general classifiers to the original model. By computing the classification scores $s = z^T w$, KEFG distinguishes novel classes with few samples and transfers the original models to other datasets efficiently.

3 Experiments

3.1 Experimental Setting

The fundamental base classes remain the training set of ImageNet 2012. We test the developmental ability on CUB and miniImageNet. The knowledge graph is

exploited from the WordNet. CUB includes 200 fine-grained classes of birds. We only take 10 classes, which are disjoint from the 1000 training classes of ImageNet 2012. MiniImageNet consists of 100 categories. For fairness, we only take 90 base classes as the training set. The remaining 10 classes in the miniImageNet consist of the novel task with few examples. The original recognition model is pre-trained on the ResNet50 [6] with base classes.

3.2 Comparision

Table 1. Comparision with prior models

Model	MiniImageNet		CUB	
Setting	1-shot	5-shot	1-shot	5-shot
Nearest neighbor [15]	44.1	55.1	52.4	66.0
Matching Nets [16]	46.6	60.0	49.3	59.3
MAML [17]	48.7	63.1	38.4	59.1
DEML+Meta-SGD [18]	58.5	71.3	66.9	77.1
Δ-encoder [14]	59.9	69.7	69.8	82.6
KEFG	61.34	77.11	73.78	84.15

The comparison between KEFG and other exiting methods is reported in Table 1, where the performance is evaluated by the average top-1 accuracy. KEFG achieves the best or competitive performance for both 10-way 1-shot and 10-way 5-shot recognition tasks. Especially, KEFG shows remarkable improvement in the fine-grained dataset. The accuracy on CUB increases almost twenty percentage the most. However, the training set in KEFG completely comes from ImageNet 2012. The relationship between the base and novel classes is weaker. We owe this excellent transferability to two aspects. First, the prestored graph knowledge provides an excellent developmental graph for the model. Second, the combination of the knowledge and the experience provides abundant information for the novel classes to refer to. Thus the transfer accuracy increases notably.

Table 2. Comparison on details

Model	Node size	Edge size	Training time
DGP [5]	32324	32538	27 min
SGCN [5]	32324	32544	20 mmin
KEFG	1010	719	7 min

In Table 2, we analyze the details. Both DGP and SGCN only exploit the knowledge graph for the inference. From the experiments, KEFG declines the

amount of subgraph a lot. Furthermore, it cuts down the training time as well. SGCN and DGP only exploit the inheritance relationship in the knowledge graph. To gather abundant information, the subgraph involved in the inference should be large. On the other hand, KEFG takes visual similarity into account, which leads to a dense graph. Thus the novel nodes can gather more information with a smaller amount of graph.

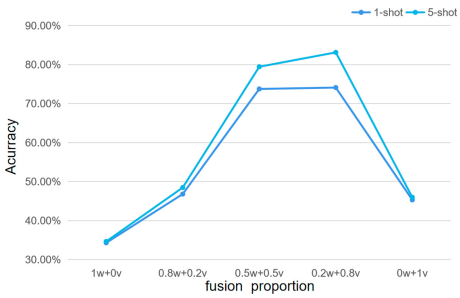

Fig. 2. In the test, w refers to the knowledge while v refers to the visual experience.

Table 3. Ablation study

Setting	1-shot	5-shot
KEFG (knowledge only)	34.33	34.67
KEFG (experience only)	45.33	48.75
KEFG (Gaussian noise+experience)	71.62	75.64
KEFG	73.78	79.47

We further test the effectiveness of the fusion idea. With different fusion proportions of knowledge and experience, the recognition accuracy changes as well. From Fig. 2, it is obviously noticed that the accuracy increases rapidly when the two sources of information are combined. After the peak accuracy of 73.78% for 1-shot and 79.47% for 5-shot, the accuracy declines as the combination becomes weak. Table 3 shows that KEFG improves the performance by almost 30% than only using knowledge or experience. Because the novel nodes gather more supplementary information from its neighbors. Both the information from word space and visual space is taken into account. Besides, not only the parent nodes and offspring nodes but also the visual similar nodes are connected to the novel ones. Furthermore, we combine the experience with Gaussian Noise to test the effectiveness of the knowledge. Table 3 shows that the combination of knowledge increases the accuracy by about 4% than Gaussian noise. Thus the knowledge information makes sense in the process of inference.

4 Conclusion

In this paper, we propose KEFG which takes advantage of information from both knowledge and experience to realize visual cognitive development. To take the interrelationship among categories into account, KEFG is based on the framework of the graph convolution network. During experiments, the ability of the proposed model outperforms previous state of art methods and obviously declines the time of training. In future work, we will devote to improving the mechanism of fusion to further improve the performance of our model.

References

1. Koch, G., Zemel, R., Salakhutdinov, R.: Siamese neural networks for one-shot image recognition. In: International Conference for Learning Representation (2017)
2. Carlson, A., Betteridge, J., Kisiel, B., Settled, B., Hruschka, E.R., Mitchell, T.M.: Toward an architecture for never ending language learning. In: AAAI (2010)
3. Wang, X.L., Ye, Y.F., Gupta, A.: Zero-shot recognition via semantic embeddings and knowledge graphs. In: CVPR (2017)
4. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: ICLR (2017)
5. Kampffmeyer, M., Chen, Y., Chen, Y.: Rethinking knowledge graph propagation for zero-shot learning. In: Conference on Computer Vision and Pattern Recognition (2019)
6. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: 2015 Machine Learning Research on International Conference for Learning Representations, vol. 15, no 1, pp. 3563–3593 (2014)
7. Miller, G.A.: WordNet: a lexical database for English. Commun. ACM **38**(11), 39–41 (1995)
8. Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al.: Matching networks for one shot learning. In: NIPS (2016)
9. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. In: NIPS (2017)
10. Finn, C., Abbeel, P., Levine, S.: Model agnostic meta-learning for fast adaptation of deep networks. In: ICML (2017)
11. Mishra, N., Rohaninejad, M., Chen, X., Abbeel, P.: A simple neural attentive meta-learner. In: ICLR (2018)
12. Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P., Hospedales, T.M.: Learning to compare: relation network for few-shot learning. In: CVPR (2018)
13. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds 200 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology (2011)
14. Schwartz, E., Kalinsky, L., Shtok, J., et al.: δ-encoder: an effective sample synthesis method for few-shot object recognition. In: NeurIPS (2018)
15. Weinberger, K.Q., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. J. Mach. Learn. Res. **10**, 207–244 (2009)
16. Ravi, S., Larochelle, H.: Optimization as a model for few-shot learning. In: International Conference on Learning Representations (ICLR), pp. 1–11 (2017)
17. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. arXiv:1703.03400 (2017)
18. Zhou, F., Wu, B., Li, Z.: Deep meta-learning: learning to learn in the concept space. arXiv:1802.03596, February 2018