Knowledge-Experience Graph with Denoising Autoencoder for Zero-Shot 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 development is vital for intelligent robots to handle various types of visual tasks rather than predefined ones. It can transfer the classification ability from an original model to a novel task. However, the high reliance on large amounts of data hinders its development. The energy it costs to adjust to the novel tasks is also a tough problem. Thus we propose a model called knowledge-experience graph (KEG) to imitate the mechanisms of human brains. With the help of social knowledge stored in the knowledge graph, the novel classes can be easily added. The combination of the experience via denoising autoencoder (DAE) also takes the relationship in the visual space into account. With the propagation of information among the graph by graph convolutional network (GCN), KEG generates the classifier of the novel tasks effectively. Experiments show that KEG improves the classification accuracy of novel categories on zero-shot learning and accomplishes visual cognitive development to a certain extent.

Keywords: GCN · Zero-shot learning · Cross-task learning · Cognitive development · Image classification · Denoising autoencoder

1 Introduction

Visual cognitive development is important for intelligent robots. With the ever-growing development of computer vision, an intelligent robot has to face various types of visual tasks rather than deterministic and predefined ones. To adjust

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

H. Yang et al. (Eds.): ICONIP 2020, CCIS 1333, pp. 121–129, 2020.
https://doi.org/10.1007/978-3-030-63823-8_15

to this unstructured and dynamic environments, a robot needs to transfer the classification ability from an original model to a novel task, while the former ability is still reserved. Cognitive development not only focuses on the cross-task problem but also deals with the zero-shot learning task. The original model has to use the unlabeled samples to retrain itself, which means it learns a novel classifier with no need of human annotation. In this way, the time and energy it takes to adjust to the novel tasks may be cut down a lot and the intelligent robot may be applied to a much tough and complicated area.

The recently proposed graph convolutional network [2] has exhibited a powerful ability in transferring knowledge across tasks. It can propagate messages among the graph and take the structural information into account. To accomplish the visual cognitive development of robots, it is reasonable to set up a neural network evolving on its own just as human brains, which is accomplished mostly by transferring information from base categories with the help of supplementary information. There are two normal sources of this information. The first one is the social knowledge developed in society, and the second one is the experience obtained based on previous tasks, which is also called empirical knowledge.

Recent researches on zero-shot learning are mostly from two viewpoints. Social knowledge builds the relation map of different classes at the macro level. Wang *et al.* [1] build an unweighted knowledge graph combined with word embedding [3,4] upon the graph convolutional network [2] to handle zero-shot problem. Kampffmeyer *et al.* [5] improve upon this model and propose Dense Graph Propagation to prevent dilution of knowledge. As for the empirical knowledge, it is acquired by recalling the related experience of the recognition task [7,8]. Gidaris *et al.* [6] get the experience as CNN is trained to recognize the base classes and propose to implement the Denoising Autoencoder network to reconstruct general weights of both the base classes and novel classes. The main part of these models is to initialize the novel categories with few samples.

Though social knowledge makes it easy to add novel classes to the map, it ignores the relationship in visual space. Empirical knowledge on the other side considers the unique visual features of the datasets. However, as the visual features are extracted from images, it can not handle zero-shot problem. Thus we argue that both these methods are not ideal for visual cognitive development.

To tackle this problem, we propose to combine social knowledge and empirical knowledge to build the relation map. The key problem for zero-shot learning is to initialize the features of novel categories with no labeled samples available. An intuitive idea is to estimates the feature of novel ones from prestored social knowledge. Based on this idea, we propose a model called knowledge-experience graph (KEG). KEG makes use of social knowledge in form of knowledge graph. The knowledge graph shows the relationship between the categories with the structure of inheritance. Novel classes aggregate supplementary information from related classes to conduct knowledge inference along the edges. Furthermore, it uses a traditional recognition model to train the base classes and observes the classification weights of base classes. Combined with the estimated value of novel classes from social knowledge, these initial weights build up an unweighted graph with the relationship of similarity. By employing the graph convolution network, information of different nodes propagates along edges and aggregates on the

novel classes iteratively. By taking the classification of base classes as ground truth, KEG finally gets the weights of novel classes and develops its cognitive ability on the novel task.

The main contributions of the paper can be summarized in three aspects. Firstly, KEG extracts social knowledge from the knowledge graph and makes it easier to add novel tasks to the original model. Secondly, based on the denoising autoencoder, the combination of the experience makes KEG focus more on the uniqueness of specific tasks. Thirdly, by introducing the graph convolutional network, the inter-cluster similarity and inter-cluster dissimilarity are taken into consideration at the same time. Thus it makes sense for KEG to deal with visual cognitive development for robots.

2 Methodology

2.1 Problem Definition

KEG focuses on visual cognitive development on the image classification task. Let C denotes all of the categories involved in the task which contains two parts *novel classes* C_{novel} and *base classes* C_{base}. The original model is trained on the C_{base} with the labeled samples, while novel classes refer to the task with no labels. According to zero-shot learning, the dataset contains two parts: the training set D_{train} with images from base classes and the testing set D_{test} with images from novel classes. Thus KEG learns from D_{train} to reconstruct a model available to D_{test} at the same time.

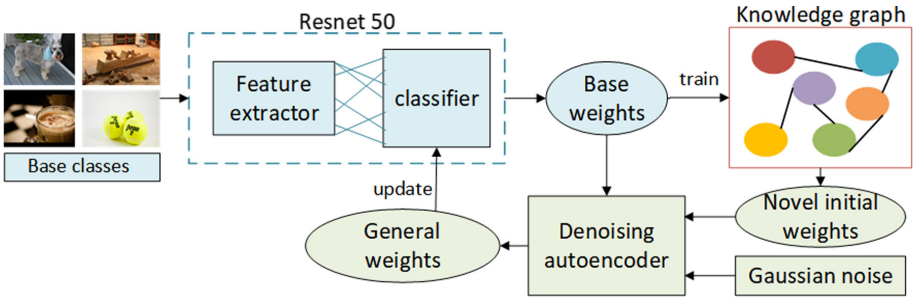

Fig. 1. Integrated framework of KEG which takes both the social knowledge and empirical knowledge into account.

2.2 Knowledge Inference Module

The knowledge graph well represents the relation map among different categories. Given an unweighted graph as $G =< V, E >$, where $V = \{v_1, v_2, ..., v_n\}$ represents the node-set of all classes, each node in it refers to a category.

$E = \{e_{i,j} = (v_i, v_j)\}$ is an edge set, if two node are related there will be an edge between them. KEG exploits the WordNet [10] as the knowledge graph to extract social knowledge. For every category, WordNet stores up its semantic description. Glove text model trained on the Wikipedia dataset is exploited to transfer the semantic description into a word embedding vector that can be operated. The feature matrix of knowledge denoted as $X_K \in R^{N \times S}$, where N is the total number of classes and S is the feature dimension of each class. For WordNet, the relationship is complicated, like hyponymy, meronymy, and troponymy. KEG builds the knowledge graph based on the hyponymy. The relationship between the nodes can be represented as

$$e_{(i,j)} = \begin{cases} 1, & \text{hyponymy(i,j)} \\ 0, & otherwise \end{cases} \tag{1}$$

The knowledge inference module works to build up the relationship among categories for zero-shot learning. The key problem is to initialize the classification weights of novel classes with no labeled samples. To gather information from related base classes to novel ones, KEG employs the graph convolutional network on the knowledge graph. For one layer of the graph neural network, a given node receives messages propagate from its neighbor along the edges and then aggregates this information combined with its status to update the class feature. The update process for a given node can be represented as

$$h^{i+1} = f(h^i, E) \tag{2}$$

where $f(x)$ refers to the mechanism of propagation and aggregation. E is the adjacent matrix and h^i is the status of the given nodes in the ith layer.

For one layer in GCN, a node only receives the information from classes connected to it. GCN can also be extended to multiple layers to perform deeper spread and get more information to perform knowledge inference. Therefore KEG employs two layer of GCN and the mechanism can be described as

$$H = \hat{D}^{-\frac{1}{2}} \hat{E} \hat{D}^{-\frac{1}{2}} ReLu(\hat{D}^{-\frac{1}{2}} \hat{E} \hat{D}^{-\frac{1}{2}} X K^{(0)}) K^{(1)} \tag{3}$$

where H denotes the output of graph, while X is the feature matrix. To reserve self information of nodes, self-loops are added among the propagation, $\hat{E} = E + I$, where $E \in R^{N \times N}$ is the symmetric adjacency matrix and $I \in R^{N \times N}$ represents identity matrix. $D_{ii} = \sum_j E_{ij}$ normalizes rows in E to prevent the scale of input modified by E. K^l is the weight matrix of the lth layer which GCN regulates constantly to achieve better performance.

During the training process, the goal is to predict the initial classification of novel classes. The graph is trained to minimize the predicted classification weights and the ground-truth weights by optimizing the loss

$$L = \frac{1}{2M} \sum_{i=1}^{M} \sum_{j=1}^{P} (W_{i,j} - W_{i,j}^k)^2, \tag{4}$$

where W^k refers to the output of base classes on GCN, which is a part of H, and W denotes the ground truth of classification weight obtained from the visual transfer model. M is the number of base classes and P is the dimensionality of the vector.

2.3 Visual Transfer Module

To take the visual feature into account, KEG learns the experience from the process the original model is trained. For an traditional classification model $C(F(\cdot|\theta)|w)$ based on CNN, it contains two parts: feature extractor $F(\cdot|\theta)$ and category classifier $C(\cdot|w^v)$ where θ and w^v indicate the parameters trained with $C_{train} = \{(\hat{x}_1, \hat{y}_1), ..., (\hat{x}_M, \hat{y}_M)\}$. $W^v \in R^{M \times P}$ refers to the classification weights that determines the classification score of each category. M is the total number of base categories and P is the length of classification weight. The goal of visual transfer module is to reconstruct a general version of classification with the framework of denoising autoencoder.

KEG also builds up a graph to represent the relationship among categories, i.e. $G^v =< X^v, E^v >$, where X^v is the node set and E^v represents the edge set. Each node refers to a category and has a visual feature X_i^v. For the base classes, the visual feature is the classification weights extracted from the original model while for the novel ones it is the initial classification from the knowledge inference model.

$$x_i^v = \begin{cases} w^v, & C_i \in C_{base} \\ w^k, & C_i \in C_{novel} \end{cases} \tag{5}$$

KEG exploits cosine similarity to generate propagation channels which are the set of edges$(i, j) \in E$ of the graph. With the boundary of cosine similarity, it can decide the density of the graph. If the visual feature of two classes are related their information can be propagated reciprocally by the edge.

$$e_{(i,j)} = \begin{cases} 1, & \frac{x_i^v \cdot x_j^v}{\|x_j^v\| \|x_j^v\|} > s \\ 0, & otherwise \end{cases} \tag{6}$$

It is worth noting that the edge is connected in terms of cosine similarity of the initial node features which is the vector before the injection of Gaussian noise. S refers to the boundary to the cosine similarity which decides the density of the graph.

To exploit the denoising autoencoder to generate the classification weights of novel classes, KEG injects Gaussian noise to the input

$$\hat{x^v} = x^v + G \tag{7}$$

G is the Gaussian noise with the same size as the node feature. Autoencoder is a neural network that generates the output by taking the input as the ground-truth. KEG uses the classification weights extracted from the original model as the ground-truth. By employing a two layers GCN on the graph, novel classes

learn the mechanism of an end to end learning of classification model from the original one and generate more universal classification weights $\tilde{W} \in R^{N \times P}$. $\tilde{W}$ is applied to the last layer of the original model which is transferred to $C(F(\cdot|\theta)|\tilde{w})$. Note that differs from W^v, $\tilde{W}$ contains n rows of P, which means it represents the classification of the whole classes C.

With the knowledge inference module and visual transfer module, KEG develops the cognitive ability to novel tasks by generating more universal classification weights. Combined with the original classification model, KEG computes the classification score of every categories as $[s_1, s_2, ..., s_N] = \{z^T\tilde{w}_1, z^T\tilde{w}_2, ..., z^T\tilde{w}_N\}$. z refers to the visual features extracted from the original model. In other words, KEG learns a mapping network, which makes a good inference from the knowledge and experience space to visual space. With the general classification scores $s = z^T\tilde{w}$, KEG distinguishes novel classes with few samples and transfers the original model to other datasets efficiently.

3 Experiment

3.1 Datasets

As KEG focus on the transfer learning of models between different datasets, ImageNet [9] is used as the base classes and AWA2 [17] as the novel classes. Besides, WordNet represents the source for constructing a knowledge graph.

ImageNet. ImageNet is an image dataset constructed base on the hierarchical structure of WordNet. We use ImageNet 2012 as the training set for zero-shot learning, which contains 1000 categories. There are no more than half of the categories are animals. Besides it also contains other classes like daily necessities, buildings, foods, which is a general dataset.

Animals with Attributes 2. AwA2 contains images of 50 animal classes with pre-extracted feature for each image. However, as we try to learn the experience from base classes, we do not use the feature it provides, but the images only. There are about ten classes that are disjoint from ImageNet and they make up the testing set in the experiment to test the transfer ability of KEG.

3.2 Experimental Setting

The original recognition model is pre-trained on ResNet50 on ImageNet 2012. The final general classification weights will adjust to the last layer of it. The output dimension of KEG is set to 2049. The model is trained in 3000 epochs. We use Adam optimizer for the training process with the weight decay of 0.0005 and the learning rate of 0.001. The boundary of similarity is set to 0.6 to ensure the density of the graph is suitable. The information of every node is mixed with both experience and knowledge equably. The whole project is under the framework of PyTorch and operated on the Ubuntu system.

3.3 Comparison

Table 1. Top-1 accuracy (%) results for classification

Model	Accuracy
SGCN [5]	74.6
SSE [11]	61.0
DEM [12]	67.1
SAE [13]	61.0
RelationNet [14]	64.2
SYNC [15]	46.6
SJE [16]	61.9
KEG	77.8

From the experiment results posted in Table 1, KEG shows better performance on zero-shot learning. It increases the classification accuracy of novel tasks. Previous methods have to extract visual features from novel classes, KEG needs no sample on novel categories. KEG stores prior social knowledge with the structure of the knowledge graph. It can easily get information from the semantic description to support its visual inference. Thus with the help of social knowledge, the way exploits empirical information expands its application range.

On the other hand, the information from the social knowledge is lack of the feature from visual space. Empirical knowledge shows the connection between categories from a visual point. From the experiment, it shows there are obvious differences in the accuracy of specific categories between KEG and SGCN, for example, 'mole'. The direct neighbors of 'mole' from the inheritance and the visual space are different. From the relationship shown in Fig. 2, we notice that besides the relationship from biology, there is also a similarity in the visual feature. For example, the dolphin belongs to the mammal but it looks more like fish. Thus it is more reasonable to gather information from the visual side since the goal of the model is to classify the image correctly.

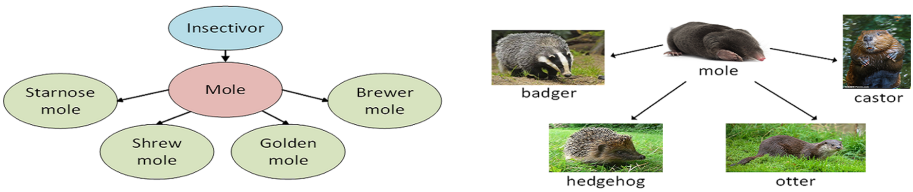

Fig. 2. The direct neighbor from the relationship of inheritance and the visual space.

3.4 Analysis of KEG

We perform ablation studies on modules of KEG to ensure that the choices we make have the best performance. Specifically, we examine the modules on the following. First, we test the similarity boundary of the connection mechanism to analyze the influence of the density of the graph. Then we use the best performance "similarity" and change the "DAE module" to ensure its importance for the increase of accuracy. The result of the ablation study is shown below.

Table 2. Top-1 accuracy results for classification

Similarity	0.8	0.6	0.5	0.4
Accuracy	76.3	77.8	73.3	73.96
DAE module	0	1		
Accuracy	74.6	77.8		

From the ablation study, we notice that a suitable similarity boundary is vital for accuracy. When the boundary is high, The similarity between categories is tight which results in a dense graph. However large boundary does not bring better performance which may be caused by dilution of information through the path. When the boundary is small, it means the relationship between the neighbor becomes further which results in a sparse graph. Since the given node can not get enough supplementary information from its neighbor the accuracy cuts down as well. Thus a suitable similarity boundary is vital for the performance. We also test the necessity of the model which shows that with the help of DAE the classification accuracy of zero-shot learning increases. The injected Gaussian noise indeed helps to reconstruct a general version of the classification weights.

4 Conclusion

In this paper, we address the problem of visual cognitive development from two parts: zero-shot learning and cross-task learning. The proposed model KEG stores social knowledge with the structure of the knowledge graph. Thus KEG builds a relation map, which supports the accession of the novel task. It also takes the feature relationship in the visual space into account with the information from the empirical knowledge. The mix of the two sources of information makes it suitable to accomplish visual cognitive development. During experiments, the ability of the proposed model outperforms previous state-of-the-art methods. In future work, we will devote to improving the mechanism of fusion to further improve the performance of our model. We also try to perform a better connection mode to avoid the attenuation of information.

References

1. Wang, X.L., Ye, Y.F., Gupta, A.: Zero-shot Recognition via Semantic Embeddings and Knowledge Graphs. In: CVPR (2017)
2. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: ICLR (2017)
3. Frome, A., et al.: Devise: a deep visual-semantic embedding model. In: Advances in Neural Information Processing Systems, pp. 2121–2129 (2013)
4. Li, Y., Wang, D., Hu, H., Lin, Y., Zhuang, Y.: Zero-shot recognition using dual visual-semantic mapping paths. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (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. Gidaris, S., Komodakis, N.: Generating classification weights with GNN Denoising Autoencoders for few-shot learning. In: Conference on Computer Vision and Pattern Recognition (2019)
7. He, K., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
8. Krizhevsky, A., Sutskever, I., Hinton, G.E.: classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.:. Imagenet: a large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009)
10. Miller, G.A.: Wordnet: a lexical database for english. Commun. ACM **38**(11), 39–41 (1995)
11. Ziming, Z., Saligrama, V.: Zero-shot learning via semantic similarity embedding. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4166–4174 (2015)
12. Zhang, L., Xiang, T., Gong, S.: Learning a deep embedding model for zero-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2021–2030 (2017)
13. Kodirov, E., Xiang, T., Gong, S.: Semantic autoencoder for zero-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3174–3183 (2017)
14. Sung, F., Yongxin, Y., Li, Z., Xiang, T., Torr, P., Hospedales, T.M.: Learning to compare: Relation network for few-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1199–1208 (2018)
15. Changpinyo, S., Wei-Lun, C., Boqing, G., Sha, F.: Synthesized classifiers for zero-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5327–5336 (2016)
16. Akata, Z., Reed, S., Walter, D., Honglak, L., Schiele, B.: Evaluation of output embeddings for fine-grained image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2927–2936 (2015)
17. Xian, Y., Lampert, C.H., Schiele, B., Akata, Z.: Zero-Shot learning - a comprehensive evaluation of the good, the bad and the ugly. IEEE Trans. Pattern Anal. Mach. Intell. **40**(8), 2251–2265 (2018)