Capsule Graph Neural Network for Multi-Label Image Recognition (Student Abstract)

Xiangping Zheng, Xun Liang, Bo Wu

School of Information, Renmin University of China, Beijing, China 100872
{xpzheng, xliang, wubochn}@ruc.edu.cn

Abstract

This paper studies the problem of learning complex relationships between multi-labels for image recognition. Its challenges come from the rich and diverse semantic information in images. However, current methods cannot fully explore the mutual interactions among labels and do not explicitly model the label co-occurrence. To overcome these shortcomings, we innovatively propose CGML that consists of two crucial modules: 1) an image representation learning module that aims to complete the feature extraction of an image whose features are expressed in the form of primary capsules; 2) a label adaptive graph convolutional network module that leverages the popular graph convolutional networks with an adaptive label correlation graph to model label dependencies. Experiments show that our approach obviously outperforms the existing state-of-the-art methods.

Introduction

Multi-label image recognition is a fundamental yet practical task in computer vision, as real-world images generally contain diverse semantic objects. Compared to single-label image classification, multi-label recognition is usually more challenging because of complex scenes, more wide label space, and implicit correlation of objects. Therefore, more effective methods are needed to identify the objects that appear in the image at the same time. In practice, some methods [Zhang and Zhou 2014] identify each object individually and naively transform this problem into multiple binary classification tasks. However, these methods ignore the topological structure between objects in the multi-label image, and the number of predicted labels will grow exponentially as the number of categories increases. For instance, keyboard, mouse, and monitor will appear in an image simultaneously with a high possibility, while keyboard and river will rarely co-occur at the same time. Therefore, a large number of label combinations are difficult to appear in the real world. Based on CNN and its variants [Huang and Liu 2020], the accuracy of existing multi-label image recognition methods is improved. However, the performance flaws of these methods are that they essentially ignore the complex topology between objects in the image. Although some previous researches [Chen, Wei, and Wang 2019; Wang et al. 2020] have shown some promising results to model label dependencies, they just focus on capturing the correlation between labels and neglect to effectively adapt among label embeddings, which inhibits the further improvement of the accuracy of multi-image recognition.

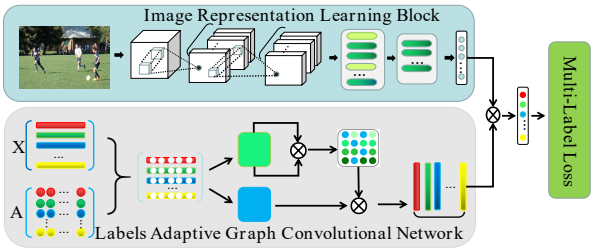

Figure 1: The framework of CGML. Our proposed model comprises two modules: IRL and LAGCN. IRL aims to complete the feature extraction of an image and LAGCN is to learn the label co-occurrence embeddings according to the relationship between different objects. Finally, we use the dot product to fuse the two-modal vectors for classification.

Methodology

In this work, we present a capsule graph neural network for multi-label image recognition named CGML to address the above challenges. The proposed method comprises two blocks: Image Representation Learning (IRL) and Label Adaptive Graph Convolutional Network (LAGCN). The motivation of this work is to effectively capture the correlations between object labels and explore these label correlations to improve classification performance. As shown in Figure 1, We now detail the proposed model as follows.

Image Representation Learning The IRL aims to complete the feature extraction of an image, and the features are expressed in the form of primary capsules. Capsule neural networks (CapsNets) have proved their effectiveness in image recognition and in a leading position by exploiting the fact that while viewpoint changes have complicated effects on pixel intensities, they have linear effects at the part/object level [Sabour, Frosst, and Hinton 2017]. In order to obtain high-quality image features, we design a new layer called the capsule selection based on the capsule idea to capture the

part-whole spatial relationships of an image. This strategy implements the selection of important lower-level capsules to ensure that only the outputs of the important capsules are sent to the layer above. The capsule selection layer not only reduces the computational burden but also improves the network generalization by collecting the outputs of the k most active capsules. Consequently, we employ global max-pooling to obtain the image-level feature $\mathbf{x}$:

$$\mathbf{x} = \mathcal{L}_{\textbf{selection}}(\mathcal{M}_{Cap}(\mathbf{I}; \theta_{Cnn})) \in \mathbb{R}^D \quad (1)$$

where $\mathbf{I}$ represents the input features of an image. θ_{Cnn} indicates model parameters.

Label Adaptive Graph Convolutional Network In this part, we use GCN to learn the label co-occurrence embeddings according to the relationship between different objects. We design the final output of each GCN node as the classifier of the corresponding label in the task. In addition to obtaining the label feature vector of each node (object), another essential issue in LAGCN is to learn the structural similarity between labels nodes. We present a label adaptive attention mechanism to update and communicate a label node's features according to its characteristics and other related neighbor nodes. To take full advantage of label long-range contextual information, we explore global contextual information by building associations among features with the label attention mechanism. Our method could adaptively aggregate long-range contextual information, thus improving feature representation for labels. Specifically, we transform the outputs of label attention modules by a convolution layer and perform an element-wise sum to accomplish feature fusion. At last, a convolution layer is followed to generate the final prediction map. Our goal is to map the object dependency of the dataset to the co-existence embedding in the marking task.

Experiments

To investigate the effectiveness of our model, we evaluate the performance of CGML and compare it with the state-of-the-art image recognition methods, including ResNet-101 [He et al. 2016], ML-GCN [Chen, Wei, and Wang 2019], A-GCN [Li et al. 2019], F-GCN [Wang et al. 2020]. We report the empirical results on two benchmark multi-label image recognition datasets: MS-COCO and VOC 2007.

The preliminary results are shown in Table 1. Obviously, CGML outperforms all candidate methods on almost all metrics. In particular, compared with three GCN-based methods (ML-GCN, A-GCN, F-GCN), we boost 3% and 2% improvement on MS-COCO and VOC 2007, respectively. The main reason is that ML-GCN adopts a graph convolution network (GCN) to capture and learn the label dependencies according to the label statistical information. However, this strategy that builds a correlation matrix by counting label pairs and thresholding is inflexible and may be sub-optimal for multi-label classification. Although this limitation can be partially alleviated by using an adaptive label graph module as reported in A-GCN. At the same time, F-GCN proposes a fast GCN to fuse image representations for multi-label image recognition. But they both focus on pursuing the relevance of labels and ignore the effective extraction of image features that are important in extracting the graph representations. In contrast, we explicitly consider the interaction between labels and effectively adapt among label embeddings. We also propose a capsule selection module to extract image features more effectively, further applied to fusing features. Therefore, we can conclude that CGML significantly promotes performance compared with other methods, which shows GCN plays a crucial role in integrating the label dependencies into image representations and demonstrates CGML has a good effect on multi-label image recognition.

MODEL	MS-COCO				VOC2007
	mAP	CF1	OF1	CF1-3	mAP
ResNet-101	77.3	71.2	75.8	69.7	89.9
ML-GCN	83.0	78.0	80.3	74.6	94.0
A-GCN	83.1	78.0	80.3	74.6	94.0
F-GCN	83.2	78.3	80.5	74.6	94.1
CGML	**86.7**	**80.5**	**81.6**	**76.8**	**96.2**

Table 1: Performance comparisons of CGML with the state-of-the-art methods (%)

Conclusions

Capturing label dependency plays one crucial issue in multi-label recognition. To better model this information, we propose a CGML framework for multi-label image recognition based on the GCN model. Experiments show that CGML can learn better feature representation for specific multi-label recognition tasks based on its label relationship.

Acknowledgments

This work was supported by the National Social Science Foundation of China (18ZDA309), the National Natural Science Foundation of China (71531012, 62072463), and the Opening Project of State Key Laboratory of Digital Publishing Technology of Founder Group (413217003). Xun Liang is the corresponding author of this paper.

References

Chen, Z.; Wei, X.; and Wang, P. 2019. Multi-Label Image Recognition With Graph Convolutional Networks. In *CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 5177–5186.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *CVPR 2016*, 770–778.

Huang, G.; and Liu, Z. 2020. Densely Connected Convolutional Networks. In *CVPR*, 2261–2269.

Li, Q.; Peng, X.; Qiao, Y.; and Peng, Q. 2019. Learning Category Correlations for Multi-label Image Recognition with Graph Networks. *CoRR*, abs/1909.13005.

Sabour, S.; Frosst, N.; and Hinton, G. E. 2017. Dynamic Routing Between Capsules. In *NIPS 2017*, 3856–3866.

Wang, Y.; Xie, Y.; Liu, Y.; Zhou, K.; and Li, X. 2020. Fast Graph Convolution Network Based Multi-label Image Recognition via Cross-modal Fusion. 1575–1584. ACM.

Zhang, M.; and Zhou, Z. 2014. A Review on Multi-Label Learning Algorithms. 26(8): 1819–1837.