Fast Graph Convolution Network Based Multi-label Image Recognition via Cross-modal Fusion

Yangtao Wang, Yanzhao Xie, Yu Liu*, Ke Zhou, Xiaocui Li
Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan, China
{ytwbruce, yzxie, liu_yu, k.zhou, LXC}@hust.edu.cn
*Corresponding author: Yu Liu (liu_yu@hust.edu.cn)

ABSTRACT

In multi-label image recognition, it has become a popular method to predict those labels that co-occur in an image via modeling the label dependencies. Previous works focus on capturing the correlation between labels, but neglect to effectively fuse the image features and label embeddings, which severely affects the convergence efficiency of the model and inhibits the further precision improvement of multi-label image recognition. To overcome this shortcoming, in this paper, we introduce Multi-modal Factorized Bilinear pooling (MFB) which works as an efficient component to fuse cross-modal embeddings and propose F-GCN, a fast graph convolution network (GCN) based multi-label image recognition model. F-GCN consists of three key modules: (1) an image representation learning module which adopts a convolution neural network (CNN) to learn and generate image representations, (2) a label co-occurrence embedding module which first obtains the label vectors via the word embeddings technique and then adopts GCN to capture label co-occurrence embeddings and (3) an MFB fusion module which efficiently fuses these cross-modal vectors to enable an end-to-end model with a multi-label loss function. We conduct extensive experiments on two multi-label datasets including MS-COCO and VOC2007. Experimental results demonstrate the MFB component efficiently fuses image representations and label co-occurrence embeddings and thus greatly improves the convergence efficiency of the model. In addition, the performance of image recognition has also been promoted compared with the state-of-the-art methods.

CCS CONCEPTS

• **Computing methodologies** → **Image representations**.

KEYWORDS

Multi-label Image Recognition; Graph Convolution Network; Cross-modal Fusion

ACM Reference Format:
Yangtao Wang, Yanzhao Xie, Yu Liu*, Ke Zhou, Xiaocui Li. 2020. Fast Graph Convolution Network Based Multi-label Image Recognition via Cross-modal Fusion. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20), October 19–23, 2020, Virtual Event, Ireland*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3340531.3411880

CIKM '20, October 19–23, 2020, Virtual Event, Ireland

ACM ISBN 978-1-4503-6859-9/20/10…$15.00
https://doi.org/10.1145/3340531.3411880

1 INTRODUCTION

In recent years, multi-label image recognition has been one of the research hotspots in computer vision community owing to its wide applications like human attribution recognition [22], music emotion categorization [30], medical diagnosis recognition [9], *etc.* Different from the conventional image classification task which only learns and predicts one label for each image, the task of multi-label image recognition brings greater challenges that call for more effective methods to recognize those objects that co-occur in an image.

Early multi-label classification algorithms [4, 31, 40] recognize each object in isolation and naively transform this problem into multiple binary classification tasks. Entering the stage of deep convolution neural network (CNN) [18], image classification has made great progress and the precision of existing multi-label image recognition methods has been promoted based on CNN and its variants [11, 14, 28]. However, the performance of these methods is essentially limited by ignoring the complex topology between objects in an image, which inhibits the further precision improvement of multi-image recognition.

An effective method to solve this problem is to model the label dependencies to learn the objective law in the real world that related objects will be more likely to co-occur in an image. As shown in Figure 1, "person", "basketball" and "basketball hoop" will appear in an image at the same time with a high possibility, while "basketball" and "mountain" will rarely co-occur in the same image. Wang *et al.* [32] utilize the recurrent neural network (RNN) to model the label dependencies in a sequential fashion, but fail to comprehensively take the correlations between image labels and image regions into consideration. To compensate for this shortcoming, some other works [33, 41] explore the label dependencies via attention mechanism, which consider limited local correlations between attended regions of a single image, but ignore the global correlations of labels distribution on all images. It is worth mentioning that Chen *et al.* [3] propose ML-GCN that adopts graph convolution network (GCN) [17, 20] to capture and learn the label dependencies according to the label statistical information, which achieves good performance. Similar to ML-GCN, Li *et al.* [21] design A-GCN to capture label dependencies by constructing an adaptive label graph. However, both of them use the dot product (DP) to simply fuse the two-modal vectors, *i.e.*, the features extracted from the CNN module and the label co-occurrence embeddings generated from the GCN module, which severely limits the convergence efficiency

of the model as well as prevents the further precision improvement of multi-label image recognition.

In this paper, we introduce Multi-modal Factorized Bilinear pooling (MFB) [39] which works as an efficient component to fuse cross-modal vectors and propose a fast GCN based model (termed as F-GCN) to fuse image representations and label embeddings for multi-label image recognition. Our F-GCN mainly contains image representation learning module, label co-occurrence embedding module and MFB fusion module. In the image representation learning module, following ML-GCN, we use a CNN (*i.e.*, ResNet-101 [11]) based model to obtain the latent representation of each image. At the same time, in the label co-occurrence embedding module, we take both the word vectors and co-occurrence correlations of labels as input, then use GCN to train and learn the label embeddings that reflect the co-occurrence relationships between different objects. Different from previous studies, in the following, we fuse these two-modal vectors (*i.e.*, image representations and label co-occurrence embeddings) via MFB to train an end-to-end multi-label image recognition model with a multi-label loss function. We conduct extensive experiments on two multi-label datasets including MS-COCO [23] and VOC2007 [6]. Experimental results demonstrate the MFB component efficiently fuses image representations and label co-occurrence embeddings and thus greatly improves the convergence efficiency of the model. In addition, the performance of image recognition has also been promoted compared with the state-of-the-art methods.

The rest of this paper is organized as follows. Section 2 talks about several related works. We formulate and introduce our F-GCN in detail in section 3. Section 4 presents the comparison experiments, ablation study as well as visual retrieval results of F-GCN. At last, we conclude this paper in section 5.

2 RELATED WORKS

In this section, we first review existing multi-label image recognition methods, then discuss recent GCN based studies, and finally introduce several representative cross-modal fusion works.

2.1 Multi-label Image Recognition

With the development of deep neural network, image recognition has achieved great success in the past few years on large scale handcrafted datasets like ImageNet [5], MS-COCO [23], *etc.* Powerful CNN based models [11, 14, 28] can extract the visual feature of each image and obtain remarkable performance for single-label classification tasks. Furthermore, researchers have made their great efforts to explore deep networks to promote the performance of multi-label image recognition.

Early multi-label image recognition methods naively divide this task into multiple independent binary classification tasks, which train a set of classifiers for each label. On the one hand, these methods suffer from the increase of labels space. Take VOC2007 as an example, there are 20 object classes in this dataset, which calls for 2^{20} classifiers if using the single-label classification method. On the other hand, they treat each object in isolation and thus neglect the topology between objects in a multi-label image. For instance, "person", "basketball" and "basketball hoop" will appear in an image at the same time with a high possibility, while "basketball" and "mountain" will rarely co-occur in the same image. Therefore, a large number of combinations of labels will hardly occur in the real world.

Various approaches have been considered to explore the label dependencies to reduce and optimize the label prediction space. Gong *et al.* [10] adopt a deep convolution architecture to learn the approximate top-k ranking objective function for multi-label image recognition. Wang *et al.* [32] combine CNN with RNN to model the label dependencies in a sequential fashion by embedding semantic labels into vectors. Besides, others propose to utilize the attention mechanism to capture the correlation between labels. Zhu *et al.* [41] propose a spatial regularization network to capture both semantic and spatial relations of these multiple labels based on weighted attention maps. Wang *et al.* [33] introduce a spatial transformer layer and long short-term memory (LSTM) units to capture the label correlation. In addition, a graph based framework [19] has been proposed to describe the relationships between labels via knowledge graphs which aims to generate more accurate image representations.

2.2 Graph Convolution Network

The basic idea of graph convolution network (GCN) [17] is to update one node's feature based on those features of the node itself and other related neighbor nodes according to the correlation matrix of the graph. By learning the structural similarities between training data points, GCN can integrate the relationships into data features. Formally, GCN takes the correlation matrix A as well as feature matrix X as input, and produces the node-level output. The forward propagation process in GCN is described as:

$$H^{l+1} = a^l(\hat{A}H^lW^l), \tag{1}$$

where $\hat{A}$ denotes the normalized version of correlation matrix A, H^l, W^l and a^l respectively denote the input, weight and non-linear activation function (like Sigmoid or ReLU) of the l-th graph convolution layer.

As a deep learning technique that effectively learns and extracts relationships, GCN has been widely applied to relational feature extraction, node classification prediction and information retrieval tasks [13, 29, 37]. Cross-model works regrade each feature of text or image as a node representation, and complete the learning process according to the mutual relationships. Jing *et al.* [38] propose to utilize GCN for text modeling and another neural network for image modeling, which achieves significant improvement with a pairwise loss function. GCH [35], another cross-modal research, learns modality-unified binary codes via an affinity graph, then adopts GCN to map hash codes by the relationships between nodes. It's worth mentioning that ML-GCN [3] can achieve remarkable performance for multi-label classification tasks. It treats each object in the image as a node, and constructs a graph among these nodes, finally uses GCN to learn the probability of different objects appearing in an image, which explores the label correlation dependency and promotes the precision of image retrieval. Similar to ML-GCN, A-GCN [21] constructs an adaptive label graph to capture label dependencies for image recognition and EmotionGCN [12] models the correlations between emotions via GCN for emotion distribution learning.

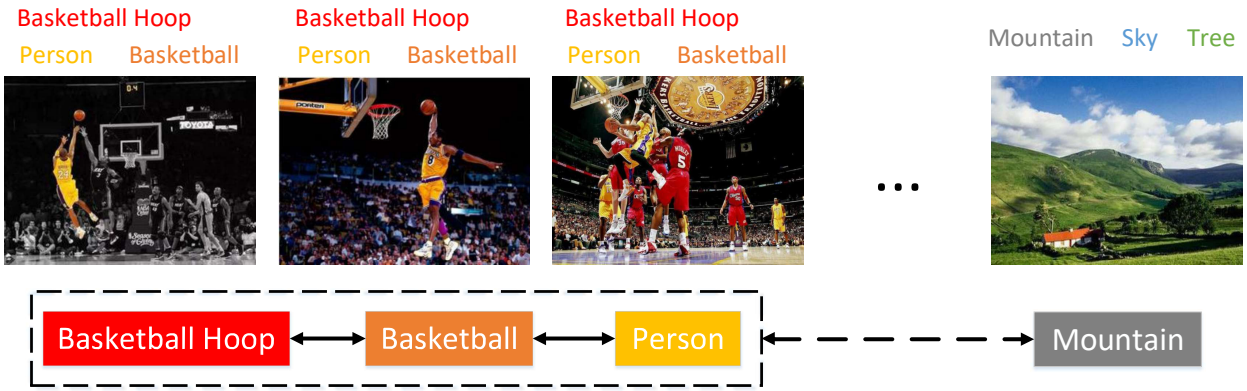


Figure 1: Related objects are more likely to appear in the same image. For example, we can always see the superstar Kobe Bryant and his teammates playing basketball on TV (in the first three images). These images are common in the real world, which illustrates the combination of "person", "basketball" and "basketball hoop" tends to co-occur at the same time. However, we hardly see a "basketball" in the "mountain", and they are rarely tied in one image, because there is no direct relationship between these two objects.

2.3 Cross-modal Fusion

Cross-modal feature fusion methods have been proposed to solve the visual question answering (VQA) problem [25], which usually use concatenation or element-wise summations to fuse the image and the question representations. Fukui *et al.* [7] first propose the Multi-modal Compact Bilinear pooling (MCB) which introduces the bilinear model to fuse multi-modal features by using the outer product of two vectors in different modalities to produce a very high-dimensional feature for quadratic expansion. To reduce the high-dimension computation, Kim *et al.* [16] propose the Multi-modal Low-rank Bilinear pooling (MLB) approach based on the Hadamard product of two feature vectors, which can achieve comparable performance to MCB but may lead to a low convergence rate. Furthermore, Zhou *et al.* [39] introduce the Multi-modal Factorized Bilinear pooling (MFB) model to efficiently fuse image and text embeddings, which produces a remarkable result in VQA as well as speeds up the model convergence.

Motivated by the above studies, our work adopts the MFB component to fuse the image representations and label co-occurrence embeddings respectively generated from the CNN and GCN modules. With the proposed F-GCN, the convergence efficiency of the model has been greatly promoted. We also demonstrate our F-GCN works as an effective model to learn the label dependencies and can be trained in an end-to-end manner with remarkable performance compared with the state-of-the-art methods.

3 PROPOSED METHODOLOGY

Based on previous studies, in this section, we propose F-GCN, a fast GCN based multi-label image recognition framework that adopts the cross-modal component (*i.e.*, MFB) to fuse image representations and label co-occurrence embeddings. Our F-GCN mainly consists of three modules: image representation learning module, label co-occurrence embedding module and MFB fusion module. In the following, we first present the overall framework of F-GCN in Figure 2, and then respectively introduce the workflow of these three modules in detail.

3.1 Overall Framework

For convenience, we list the preliminary notations in Table 1. As shown in Figure 2, there are three key modules in F-GCN: a CNN module for image feature extraction, a GCN module for co-occurrence label embedding generation and an MFB fusion module for cross-modal vectors fusion.

Given a dataset consisting of N images, the i-th image x_i is input to a CNN module (*i.e.*, ResNet-101 [11]) to extract its image representation f_i (a D-dimension feature vector) from the "*conv5_x*"

Table 1: Preliminary notations used in this paper.

Notation	Explanation
N	the number of input images
x_i	the i-th image
f_i	the i-th image's representation
l_i	the i-th image's ground truth label
y_i	the i-th image's predicted label
C	the number of object categories
o_i	the i-th object in the label set
T_i	the occurrence times of the i-th object
T_{ij}	the co-occurrence times of the i-th and j-th objects
d	the dimension of each object's word embedding vector
D	the dimension of each image's representation
Z	a $C \times d$ object word embeddings matrix
A	a $C \times C$ label correlation matrix
W	a $C \times D$ label co-occurrence embeddings matrix
W_j	the j-th row vector of W
U	the weights in GCN propagation
M	the dimension of M_1 (or M_2) in MFB fusion
g	the number of units in each *group sum-pooling*

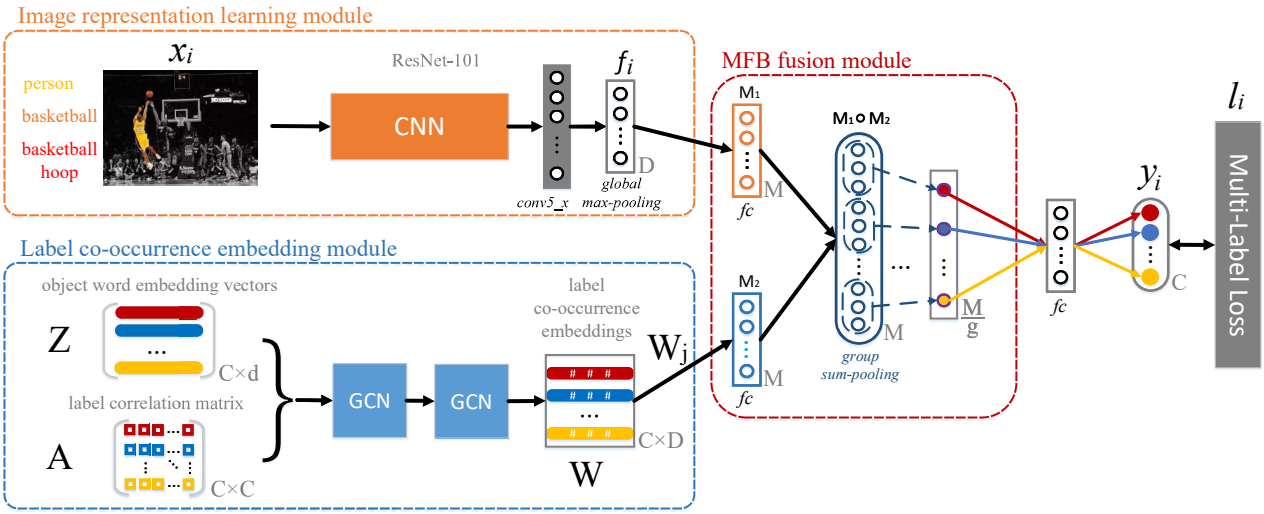


Figure 2: The overall framework of our proposed F-GCN is comprised of three key modules: image representation learning module, label co-occurrence embedding module and MFB fusion module. The image representation module uses a CNN backbone (*i.e.*, ResNet-101) to train and extract the representation (feature) of each image. The label co-occurrence embedding module designs a two-layer GCN to learn the co-occurrence embeddings that reflect the label dependencies according to the label statistical information. The MFB module efficiently fuses these two cross-modal vectors by means of *group sum-pooling*. The overall network is trained in an end-to-end manner using the Multi-Label Loss function.

layer of this network. At the same time, the GCN module takes both the object word embeddings matrix Z and label correlation matrix A as input, then adopts a two-layer GCN to learn and capture the label co-occurrence embeddings matrix W. After obtaining the image representations and label co-occurrence embeddings, we design the MFB fusion module to efficiently fuse these two cross-modal vectors (f_i and W) via *group sum-pooling*, and use the fully connected (fc) layer to generate the predicted label y_i. Finally, the multi-label loss function is adopted to update the loss between y_i and the ground truth label $l_i \in \{0, 1\}^C$ to train the whole network in an end-to-end manner.

3.2 Image Representation Learning Module

Following ML-GCN [3], we adopt one of the state-of-the-art CNN based networks (*i.e.*, ResNet-101 [11]) to complete the feature extraction of an image in this module. As shown in the orange frame of Figure 2, we first remove the last fc layer and $softmax$ layer of ResNet-101, and then use this sub-network to generate the representation of each image x_i with D dimension. As we know, for ResNet-101, the output dimension D is 2048. Therefore, for any an input image x_i with the 448×448 resolution, we can obtain $2048 \times 14 \times 14$ feature maps from the "$conv5\_x$" layer. At last, we adopt *global max-pooling* to generate the image representation f_i which is a 2048-dimension feature vector.

3.3 Label Co-occurrence Embedding Module

In this part, we use GCN to learn the label co-occurrence embeddings according to the relationship between different objects.

Different from the original GCN which was proposed to solve the node classification problem, we treat and design the node-level output of our F-GCN as a classifier corresponding to each label. Specifically, we aim to map the object dependencies of a dataset to label co-occurrence embeddings in our task. The input of GCN calls for the feature vector of each node and the correlation matrix between these nodes. As shown in the blue frame of Figure 2, we adopt the GloVe [27] model to transform each object (totally C objects in a dataset) into a d-dimensional (*i.e.*, 300-dimensional) word vector. Therefore, we can obtain a $C \times d$ object word embeddings matrix Z. For example, there are 20 object categories in VOC2007, so the input feature vectors matrix Z for the first GCN layer will be a 20×300 matrix.

In addition to obtaining the feature vector of each node (object), another essential issue in F-GCN is to construct the label correlation matrix A between these nodes. In the implementation, we capture the label dependencies and construct matrix A according to the label statistical information over the whole dataset. Specifically, for $\forall i \in [1, C]$, we collect the occurrence times (*i.e.*, T_i) of the i-th object (*i.e.*, o_i) as well as the co-occurrence times (*i.e.*, T_{ij}, which equals T_{ji}) of o_i and o_j. Furthermore, the label dependencies can be formulated by the conditional probability as follows:

$$P_{ij} = P(o_i|o_j) = \frac{T_{ij}}{T_j}, \tag{2}$$

where P_{ij} denotes the probability that o_i occurs in the conditional of o_j appearing. Note that P_{ij} is not equal to P_{ji} owing that the conditional probability between two objects is asymmetric. Based on this, we can construct the correlation matrix A below:

$$A_{ij} = P_{ij}, \tag{3}$$

where A_{ij} denotes the i-th row and j-th column element of matrix A. However, similar to ML-GCN, if we directly use this correlation matrix to train the model, the rare co-occurrence objects will become some noise that affect the data distribution as well as the model convergence. To filter the noise, we choose to use a threshold ϵ to binarize the above matrix A:

$$A_{ij} = \begin{cases} 0, & \text{if } P_{ji} \leq \epsilon \\ 1, & \text{otherwise} \end{cases} \tag{4}$$

where $\epsilon \in [0, 1]$. Besides, when using GCN to update the node's feature in the propagation process, the binary correlation matrix may lead to the over-smoothing problem which makes the generated nodes' features indistinguishable. Therefore, we adopt the weighted scheme to calculate the final correlation matrix as:

$$A_{ij} = \begin{cases} \frac{\delta}{\sum_{j=1 \cap i \neq j}^{C}} A_{ij}, & \text{if } i \neq j \\ 1 - \delta, & \text{otherwise} \end{cases} \tag{5}$$

where $\delta \in [0, 1]$. In this way, we can use this weighted correlation matrix A to update the node's feature by choosing a suitable δ.

After obtaining both the object word embeddings vectors Z and label correlation matrix A, we design a two-layer GCN to propagate this relationship and each GCN layer can be described as:

$$Z^{l+1} = f^l(\hat{A}Z^l U^l), \quad l \in [0, 2] \tag{6}$$

where $\hat{A}$ (see [17] for details) denotes the normalized version of correlation matrix A, Z^l denotes the latent features of C nodes in the l-layer, U^l denotes the weights of the l-layer, and $f^l(\cdot)$ denotes the non-linear operation which is a ReLU function. Note that Z is the input of this sub-network, and the output is a $C \times D$ label co-occurrence embeddings matrix W of which each row will be fused with the image representation f_i in the next MFB fusion module.

3.4 MFB Fusion Module

MFB [39] has been proposed to work as an effective component to fuse cross-modal features, which is usually implemented by combining some commonly-used layers such as fc, element-wise multiplication and pooling layers. Different from the previous works that use DP to simply fuse the cross-modal vectors, in this part, we adopt MFB to efficiently fuse image representations and label co-occurrence embeddings, which helps achieve higher performance of F-GCN.

As shown in the red frame of Figure 2, on the one hand, the Hadmard product increases the interaction between different modal vectors, which promotes the precision of F-GCN. On the other hand, *group sum-pooling* reduces over-fitting and parameters explosion, which speeds up the convergence of F-GCN. The input of MFB consists of two parts: f_i and W. Note that in each fusion, we use f_i and one row vector of W to generate one of the element in y_i.

Formally, given the i-th image representation f_i, for $\forall j \in [1, C]$, we fuse f_i and W_j to generate the j-th element of y_i where Wj is stated to be the j-th row vector of W in Table 1. First, we respectively use two fc layers to transform f_i to a M-dimensional vector M_1 and W_j to a M-dimensional vector M_2. Then these two modal vectors are fused to generate a M-dimensional vector $M_1 \circ M_2$, where $\circ$ is the Hadamard product. To speed up the convergence, we use *group sum-pooling* to convert $M_1 \circ M_2$ to a $\frac{M}{g}$-dimensional vector, where each group containing g units is sequentially mapped into one unit. Finally, we adopt a fc layer to generate the j-th element of y_i. We will obtain a complete predicted label vector y_i corresponding to f_i after C times fusion with f_i. Note that all the fc layers are shared by each fusion.

Finally, we adopt the MultiLabelSoftMarginLoss[1] (termed as Multi-label loss) function to update the whole network in an end-to-end manner. The training loss function is described as:

$$\mathcal{L}(y_i, l_i) = -\frac{1}{C}\sum_{j=1}^{C} l_{ij} log((1 + exp(-y_{ij}))^{-1}) + (1 - l_{ij}) log(\frac{exp(-y_{ij})}{(1+exp(-y_{ij}))}), \tag{7}$$

where y_{ij} and l_{ij} respectively denote the j-th element of y_i and l_i.

4 EXPERIMENTS

In this section, we evaluate the performance of F-GCN and compare it with the state-of-the-art image recognition methods. We first describe the datasets, then introduce the implementation details and evaluation metrics, and finally present the experimental results of F-GCN.

4.1 Datasets

MS-COCO [23] is a popular multi-label dataset for image recognition, segmentation and captioning, which contains 118,287 training images, 40,504 validation images and 40,775 test images, where each image is averagely labeled with about 2.9 object labels from the 80 semantic class categories except that the labels of test set are not available. Owing that the ground truth labels of the test set are not available, we train our model on the train set and evaluate the performance on the validation set.

VOC2007 [6] consists of 9,963 multi-label images and 20 object classes, which is divided into train, validation and test sets. On average, each image is annotated with 1.5 labels. We use both the train and validation sets to train our model, and then evaluate the performance on the test set.

4.2 Implementation Details and Evaluation Metrics

Implementation details. All experiments are implemented with PyTorch. In the image representation module, each image is resized into 448×448 using random horizontal flips and the output dimension is $D = 2048$ from ResNet-101. In the label co-occurrence embedding module, our F-GCN consists of a two-layer GCN with output dimension of 1024 and 2048, where the initial label word embedding is a 300-dimensional vector generated by the GloVe model pre-trained on the Wikipedia dataset. Note that we use the average embeddings of all words as the label word vector if this label is expressed by multiple words. To construct the correlation matrix, we respectively set $\epsilon = 0.4$ in Equation 4 and $\delta = 0.2$ in Equation 5. In the MFB fusion module, we set $M = 358$ to fuse cross-modal embeddings and $g = 2$ to complete *group sum-pooling*. The whole network is updated by stochastic gradient descent (SGD) with a momentum of 0.9, a weight decay of 10^{-4}, an initial learning

[1] https://pytorch.org/docs/master/nn.html?highlight=multilabelsoft#torch.nn.MultiLabelSoftMarginLoss

Table 2: Performance comparisons of F-GCN with the state-of-the-art methods on MS-COCO.

Method	All							Top-3					
	mAP	CP	CR	CF1	OP	OR	OF1	CP	CR	CF1	OP	OR	OF1
CNN-RNN [32]	61.2	-	-	-	-	-	-	66.0	55.6	60.4	69.2	66.4	67.8
RNN-Attention [33]	-	-	-	-	-	-	-	79.1	58.7	67.4	84.0	63.0	72.0
Order-Free RNN [1]	-	-	-	-	-	-	-	71.6	54.8	62.1	74.2	62.2	67.7
ML-ZSL [19]	-	-	-	-	-	-	-	74.1	**64.5**	69.0	-	-	-
SRN [41]	77.1	81.6	65.4	71.2	82.7	69.9	75.8	85.2	58.8	67.4	87.4	62.5	72.9
Multi-Evidence [8]	-	80.4	70.2	74.9	85.2	72.5	78.4	84.5	62.2	70.6	89.1	64.3	74.7
ResNet-101 [11]	77.3	80.2	66.7	72.8	83.9	70.8	76.8	84.1	59.4	69.7	89.1	62.8	73.6
ML-GCN [3] (DP)	83.0	85.1	72.0	78.0	85.8	75.4	80.3	89.2	64.1	74.6	**90.5**	66.5	**76.7**
A-GCN [21] (DP)	83.1	84.7	72.3	78.0	85.6	75.5	80.3	89.0	64.2	74.6	**90.5**	66.3	76.6
F-GCN (MFB)	**83.2**	**85.4**	**72.4**	**78.3**	**86.0**	**75.7**	**80.5**	**89.3**	64.3	**74.7**	**90.5**	**66.6**	**76.7**

Table 3: AP and mAP comparisons of F-GCN with the state-of-the-art methods on VOC2007.

Method	AP																				mAP
	areo	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	
CNN-RNN [32]	96.7	83.1	94.2	92.8	61.2	82.1	89.1	94.2	64.2	83.6	70.0	92.4	91.7	84.2	93.7	59.8	93.2	75.3	**99.7**	78.6	84.0
RLSD [24]	96.4	92.7	93.8	94.1	71.2	92.5	94.2	95.7	74.3	90.0	74.2	95.4	96.2	92.1	97.9	66.9	93.5	73.7	97.5	87.6	88.5
VeryDeep [28]	98.9	95.0	96.8	95.4	69.7	90.4	93.5	96.0	74.2	86.6	**87.8**	96.0	96.3	93.1	97.2	70.0	92.1	80.3	98.1	87.0	89.7
ResNet-101 [11]	**99.5**	97.7	97.8	96.4	65.7	91.8	96.1	97.6	74.2	80.9	85.0	**98.4**	96.5	95.9	98.4	70.1	88.3	80.2	98.9	89.2	89.9
FeV+LV [36]	97.9	97.0	96.6	94.6	73.6	93.9	96.5	95.5	73.7	90.3	82.0	95.4	97.7	95.9	98.6	77.6	88.7	78.0	98.3	89.0	90.6
HCP [34]	98.6	97.1	98.0	95.6	75.3	94.7	95.8	97.3	73.1	90.2	80.0	97.4	96.1	94.9	96.3	78.3	94.7	76.2	97.9	91.5	90.9
RNN-Attention [33]	98.6	97.4	96.3	96.2	75.2	92.4	96.5	97.1	76.5	92.0	87.7	96.8	97.5	93.8	98.5	81.6	93.7	82.8	98.6	89.3	91.9
Atten-Reinforce [2]	98.6	97.1	97.1	95.5	75.6	92.8	96.8	97.3	78.3	92.2	87.6	96.9	96.5	93.6	98.5	81.6	93.1	83.2	98.5	89.3	92.0
ML-GCN [3] (DP)	**99.5**	**98.5**	98.6	98.1	80.8	94.6	97.2	98.2	82.3	**95.7**	86.4	98.2	**98.4**	**96.7**	**99.0**	84.7	**96.7**	84.3	98.9	**93.7**	94.0
A-GCN [21] (DP)	99.4	**98.5**	98.6	98.0	80.8	94.7	97.2	98.2	82.4	95.5	86.4	98.2	**98.4**	**96.7**	98.9	**84.8**	96.6	**84.4**	98.9	**93.7**	94.0
F-GCN (MFB)	**99.5**	**98.5**	**98.7**	**98.2**	**80.9**	**94.8**	**97.3**	**98.3**	**82.5**	**95.7**	86.6	98.2	**98.4**	**96.7**	**99.0**	**84.8**	**96.7**	**84.4**	99.0	**93.7**	**94.1**

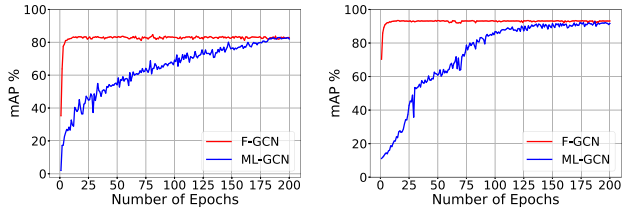
rate of 0.1 which decays by a factor of 10 every 40 epochs, and a batchsize of 32.

Evaluation metrics. We use the conventional image recognition evaluation metrics including the mean of class-average precision (mAP), overall precision (OP), recall (OR), F1 (OF1), and average per-class precision (CP), recall (CR), F1 (CF1). For each image, the labels are predicted as positive if the confidences of them are greater than 0.5. In addition, we also present these evaluation results of top-3 labels for fair comparisons.

4.3 Experimental Results

In this part, we first show the convergence efficiency of F-CGN, then compare it with the state-of-the-art image recognition methods as well as conduct ablation study to explore the influence of different parameters and components on the model, and finally give the visual retrieval results.

4.3.1 ***Convergence efficiency.*** We compare the convergence efficiency of F-GCN with that of ML-GCN. For fair comparisons, we employ the same training parameters (SGD, learning rate, batchsize, *etc*), loss function (Multi-label loss function) and datasets (MS-COCO and VOC2007) as ML-GCN. As shown in Figure 3, we show the convergence trend of mAP on the test set when increasing the epochs on the training set. As we see, F-GCN has converged at the 17-th and 15-th epoch on MS-COCO and VOC2007 respectively, and produces the higher mAP of 83.2% and 94.1%. However, at the same time, ML-GCN never converges and its mAP values are respectively 38.89% and 72.91% lower than our F-GCN. In addition, ML-GCN will take about 200 epochs (more than 11 times of F-GCN) to complete its training process. This result verifies that our MFB component efficiently fuses the cross-modal embeddings and greatly promotes the convergence efficiency of F-GCN.



(a) MS-COCO (b) VOC2007

Figure 3: mAP on test set with the increase of epoch on training set.

4.3.2 ***Comparisons with the state-of-the-art methods.*** In this part, we respectively conduct experiments on MS-COCO and VOC2007 to compare the performance of F-GCN with the the state-of-the-art methods. Note that, for fair comparisons, we implement our F-GCN using the same feature extraction network (*i.e.*, ResNet-101) and weighted correlation matrix as ML-GCN.

Results on MS-COCO. We compare F-GCN with the state-of-the-art methods including CNN-RNN [32], RNN-Attention [33], Order-Free RNN [1], ML-ZSL [19], SRN [41], Multi-Evidence [8], ResNet-101 [11], ML-GCN [3] and A-GCN [21]. We list the comparison results on MS-COCO in Table 2 including the evaluation metrics over the whole dataset and the top-3 labels. Obviously, F-GCN outperforms all candidate methods on almost all metrics. Specifically, F-GCN greatly promotes the performance compared with ResNet-101 baseline, which shows GCN plays a crucial role in integrating the label dependencies into image representations. In addition, compared with ML-GCN and A-GCN that use DP to fuse images representations and label co-occurrence embeddings, our F-GCN further improve mAP and other metrics results, which verifies that MFB in our framework can effectively fuse cross-modal embeddings.

Results on VOC2007. We compare F-GCN with the state-of-the-art methods including CNN-RNN [32], RLSD [24], VeryDeep [28], ResNet-101 [11], FeV+LV [36], HCP [34], RNN-Attention [33], Atten-Reinforce [2], ML-GCN [3] and A-GCN [21]. We list the AP and mAP results on VOC2007 in Table 3. Similarly, our F-GCN greatly outperforms the baseline (ResNet-101) and other well-known methods in all metrics except for a lower result on "table", "dog" and "train" objects. Note that we generate a higher mAP than both ML-GCN and A-GCN, which demonstrates F-GCN has an good effect on multi-label image recognition.

In general, according to the comparison results, F-GCN not only greatly speeds up the convergence efficiency but also promotes the performance of multi-label image recognition via cross-modal fusion.

4.3.3 ***Ablation study.*** In this section, we conduct ablation study to explore the influence of different parameters and components on our model including image representation extraction model, word embedding methods for label vectors, parameters ϵ, δ in correlation matrix, different number of layers in GCN, the dimension M in cross-modal fusion and the number of units g in *group sum-pooling*. Note that we use mAP, CF1, OF1, CF1-3 and OF1-3 as the evaluation metrics of our F-GCN.

Image representation extraction model. In this part, we evaluate the performance of F-GCN by comparing two commonly-used feature extraction CNN based models: VGG [28] and ResNet-101 [11]. We list the results on MS-COCO (all labels and top-3 labels) and VOC2007 in Table 4. As we see, ResNet-101 produces a higher performance than VGG. This may lies in that the later ResNet-101 has a strong ability to extract features which will be fused with the label co-occurrence embeddings to generate a better model.

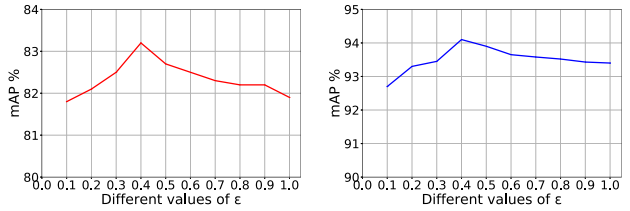
Word embedding methods for label vectors. In this part, we evaluate the performance of F-GCN by trying several popular word embedding methods including GloVe [27], GoogleNews [26], FastText [15] and the simple one-hot encoding technique to generate label vectors. We list the experimental results on MS-COCO and VOC2007 in Table 5. As we see, different word embedding methods have a slight impact on the results of F-GCN, except that GloVe gains a small preponderance than others. This illustrates the input correlations of GCN with not severely affect the result, but it is GCN that plays a crucial role in propagating and capturing the label dependencies. Of course, we believe more powerful word embeddings can maintain the semantic topology between objects and our F-GCN may benefit from a better scheme. Given this, we choose GloVe to obtain the label vectors by default.

Table 4: Results of different image feature extraction models.

Model	Dataset					
	MS-COCO					VOC2007
	mAP	CF1	OF1	CF1-3	OF1-3	mAP
VGG	82.1	77.1	78	73.5	75.7	92.9
ResNet-101	**83.2**	**78.3**	**80.5**	**74.7**	**76.7**	**94.1**

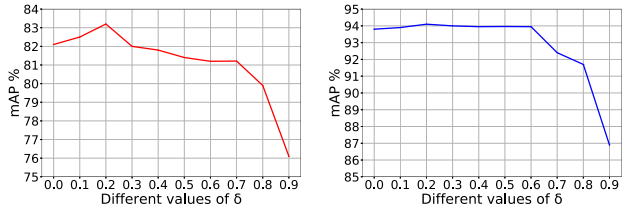
Table 5: Results of different word embedding methods (WEM) for label vectors.

WEM	Dataset					
	MS-COCO					VOC2007
	mAP	CF1	OF1	CF1-3	OF1-3	mAP
GloVe	**83.2**	**78.3**	**80.5**	**74.7**	**76.7**	**94.1**
GoogleNews	83.1	78.2	80.4	74.5	**76.7**	94.0
FastText	**83.2**	78.2	80.3	74.5	76.6	94.0
OneHot	83.1	78.2	80.3	75.5	76.6	94.0



(a) MS-COCO (b) VOC2007

Figure 4: The change of mAP using different values of ϵ.



(a) MS-COCO (b) VOC2007

Figure 5: The change of mAP using different values of δ.

Parameters ϵ and δ in correlation matrix. In this part, we evaluate the performance of F-GCN by using different values of ϵ (Equation 4) and δ (Equation 5) to construct the weighted correlation matrix.

Parameter ϵ is used to filter the noise data (rare co-occurrence probability) to enable a better model. As shown in Figure 4, we vary ϵ from 0.1 to 1 to observe the effect and find that F-GCN achieves the highest mAP on both MS-COCO and VOC2007 when $\epsilon = 0.4$. This may result from that $\epsilon = 0.4$ is a better balance parameter which not only reduces the small-probability data points but also reserves the correlation between objects.

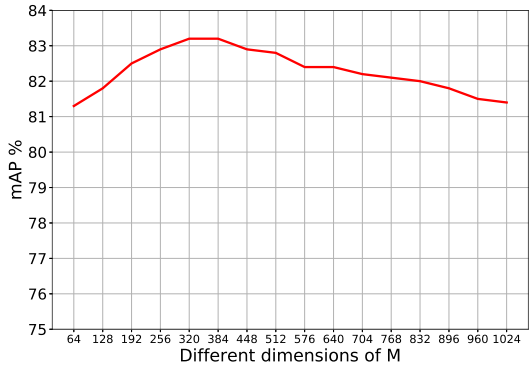
In addition, we also explore the influence of parameter δ. As we mentioned, δ is used to avoid the over-smoothing problem which may make the generated nodes' features indistinguishable. We vary δ from 0 to 1 to observe the effect and find F-GCN achieves the highest mAP on both MS-COCO and VOC2007 when $\delta = 0.2$. Note that F-GCN will not converge when setting $\delta = 1$. If we use a larger δ, the feature of the node itself may be ignored in the propagation process. Otherwise, a too smaller δ will make F-GCN ignore the correlation between a node with its neighbor nodes. The result demonstrates $\delta = 0.2$ can well balance this correlation.

Different number of layers in GCN. In this part, we evaluate the performance of F-GCN by changing the number of layers in GCN. We respectively use 2-layer (with the output dimension of 1024 and 2048), 3-layer (with the output dimension of 1024, 1024, and 2048) and 4-layer (with the output dimension of 1024, 1024, 1024 and 2048) GCN to train our model. We list the comparison results in Table 6. As we see, with the increase of layers, the performance begins to decrease on both MS-COCO and VOC2007. This may result from that in the propagation process, the features of nodes will be frequently accumulated with more GCN layers so that the output features become indistinguishable, thereby reducing the performance of image recognition.

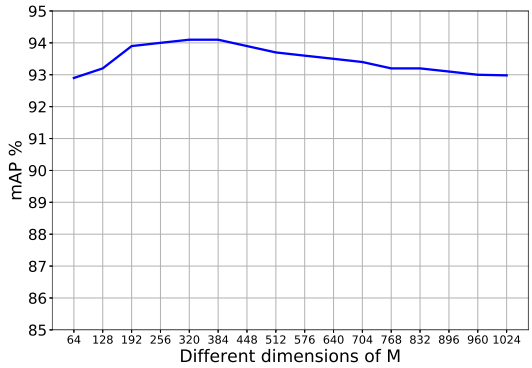
Table 6: Different number of layers in GCN.

# layers	Dataset					
	MS-COCO					VOC2007
	mAP	CF1	OF1	CF1-3	OF1-3	mAP
2 layers	**83.2**	**78.3**	**80.5**	**74.7**	**76.7**	**94.1**
3 layers	82.4	76.9	79.8	73.7	76.3	93.7
4 layers	82.3	76.7	79.6	73.1	75.9	93.2

The dimension M in cross-modal fusion. In this part, we evaluate the performance of F-GCN by changing the dimension of M when fusing the image representations and label co-occurrence embeddings. The input of MFB fusion module consists of 2048-dimensional vectors pairs, which will be reduced into M-dimension via fc layers. We vary M from 64 to 1024 with the step size of 64. As shown in Figure 6, the performance of F-GCN will be improved with the increase of M until M exceeds 384 on MS-COCO and VOC2007. Maybe $M \in [320, 384]$ not only plays a better role in dimensionality reduction, but also efficiently fuses the cross-modal vectors. In fact, in the experiment, we find $M = 358$ can bring a good result for both the efficiency and precision. We believe a more detailed perspective about the effect of M will be given if the interval is divided more finely.

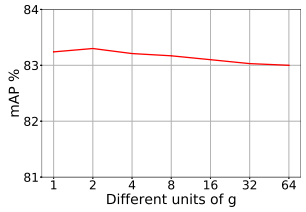


(a) MS-COCO

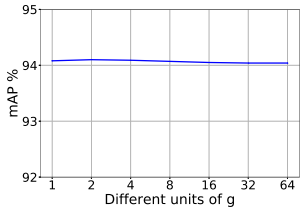


(b) VOC2007

Figure 6: The change of mAP using different dimension of M in cross-modal fusion.



(a) MS-COCO



(b) VOC2007

Figure 7: The change of mAP using different units of g.

The number of units g in *group sum-pooling*. In this part, we evaluate the performance of F-GCN by using different number of units g. By *group sum-pooling*, each M-dimensional vector will be transformed into a $\frac{M}{g}$-dimensional vector. We vary the value of g from 1 to 64 to generate a light-weight fusion vector. As shown in Figure 7(a), F-GCN obtains a better performance on MS-COCO when choosing $g = 2$, while the change of mAP is very slight on VOC2007 in Figure 7(b). We believe $g = 2$ can better express the original semantic information by pooling. Otherwise, other values of g also bring a comparable result, which will not indeed affect the model too much. It is the structure of MFB that plays a vital role in promoting the performance of F-GCN.

Figure 8: Two examples of the returned results with the query image on VOC2007.

4.3.4 ***Visual retrieval results.*** In this section, we evaluate F-GCN by giving two retrieval examples on VOC2007. We return the top-5 images by the k-NN algorithm for each given query image. Figure 8 lists the retrieval results. For example, the first input image contains two objects: "person" and "dog", and each returned image also contains these two objects. Besides, we obtain a similar effect when inputting the second image that contains "bus" and "car". The visual retrieval results verify that F-GCN owns a good classification ability to recognize multi-label images.

5 CONCLUSION AND FUTURE WORK

In order to model the label dependencies and efficiently fuse cross-modal vectors (*i.e.*, image representations and label co-occurrence embeddings), in this paper, we introduce a cross-modal fusion component (*i.e.*, MFB) and propose F-GCN, a fast GCN based multi-label image recognition model. F-GCN first respectively adopts a CNN to extract the image features and a GCN to capture the label co-occurrence embeddings according to the relationship between different objects, then utilizes MFB to efficiently fuse these cross-modal embeddings and trains an end-to-end model with a multi-label loss function. Extensive experimental results on MS-COCO and VOC2007 demonstrate the MFB component efficiently fuses image representations and label co-occurrence embeddings and thus greatly improves the convergence efficiency of the model. In addition, the performance of image recognition has also been promoted compared with the state-of-the-art methods. In the future, we will integrate the attention mechanism into our model to extract more accurate image features to help further promote the image recognition performance.

ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China No.61902135 and the Innovation Group Project of the National Natural Science Foundation of China No.61821003. Thanks for Jay Chou, a celebrated Chinese singer whose songs have been accompanying the author.

REFERENCES

[1] Shang-Fu Chen, Yi-Chen Chen, Chih-Kuan Yeh, and Yu-Chiang Frank Wang. 2018. Order-Free RNN With Visual Attention for Multi-Label Classification. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, Sheila A. McIlraith and Kilian Q. Weinberger (Eds.). AAAI Press, 6714–6721. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16114

[2] Tianshui Chen, Zhouxia Wang, Guanbin Li, and Liang Lin. 2018. Recurrent Attentional Reinforcement Learning for Multi-Label Image Recognition. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, Sheila A. McIlraith and Kilian Q. Weinberger (Eds.). AAAI Press, 6730–6737. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16654

[3] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. 2019. Multi-Label Image Recognition With Graph Convolutional Networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 5177–5186. DOI: http://dx.doi.org/10.1109/CVPR.2019.00532

[4] Amanda Clare and Ross D. King. 2001. Knowledge Discovery in Multi-label Phenotype Data. In *Principles of Data Mining and Knowledge Discovery, 5th European Conference, PKDD 2001, Freiburg, Germany, September 3-5, 2001, Proceedings (Lecture Notes in Computer Science)*, Luc De Raedt and Arno Siebes (Eds.), Vol. 2168. Springer, 42–53. DOI: http://dx.doi.org/10.1007/3-540-44794-6_4

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*. IEEE Computer Society, 248–255. DOI: http://dx.doi.org/10.1109/CVPR.2009.5206848

[6] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. 2010. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision* 88, 2 (2010), 303–338. DOI: http://dx.doi.org/10.1007/s11263-009-0275-4

[7] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, Jian Su, Xavier Carreras, and Kevin Duh (Eds.). The Association for Computational Linguistics, 457–468. DOI: http://dx.doi.org/10.18653/v1/d16-1044

[8] Weifeng Ge, Sibei Yang, and Yizhou Yu. 2018. Multi-Evidence Filtering and Fusion for Multi-Label Classification, Object Detection and Semantic Segmentation Based on Weakly Supervised Learning. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 1277–1286. DOI: http://dx.doi.org/10.1109/CVPR.2018.00139

[9] Zongyuan Ge, Dwarikanath Mahapatra, Suman Sedai, Rahil Garnavi, and Rajib Chakravorty. 2018. Chest X-rays Classification: A Multi-Label and Fine-Grained Problem. *CoRR* abs/1807.07247 (2018). arXiv:1807.07247 http://arxiv.org/abs/1807.07247

[10] Yunchao Gong, Yangqing Jia, Thomas Leung, Alexander Toshev, and Sergey Ioffe. 2014. Deep Convolutional Ranking for Multilabel Image Annotation. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1312.4894

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 770–778. DOI: http://dx.doi.org/10.1109/CVPR.2016.90

[12] Tao He and Xiaoming Jin. 2019. Image Emotion Distribution Learning with Graph Convolutional Networks. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval, ICMR 2019, Ottawa, ON, Canada, June 10-13, 2019*, Abdulmotaleb El-Saddik, Alberto Del Bimbo, Zhongfei Zhang, Alexander G. Hauptmann, K. Selçuk Candan, Marco Bertini, Lexing Xie, and Xiao-Yong Wei (Eds.). ACM, 382–390. DOI: http://dx.doi.org/10.1145/3323873.3326593

[13] Fenyu Hu, Yanqiao Zhu, Shu Wu, Liang Wang, and Tieniu Tan. 2019. Hierarchical Graph Convolutional Networks for Semi-supervised Node Classification. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, Sarit Kraus (Ed.). ijcai.org, 4532–4539. DOI: http://dx.doi.org/10.24963/ijcai.2019/630

[14] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2017. Densely Connected Convolutional Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2261–2269. DOI: http://dx.doi.org/10.1109/CVPR.2017.243

[15] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hervé Jégou, and Tomas Mikolov. 2016. FastText.zip: Compressing text classification models. *CoRR* abs/1612.03651 (2016). arXiv:1612.03651 http://arxiv.org/abs/1612.03651

[16] Jin-Hwa Kim, Kyoung Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. 2017. Hadamard Product for Low-rank Bilinear Pooling. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net. https://openreview.net/forum?id=r1rhWnZkg

[17] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net. https://openreview.net/forum?id=SJU4ayYgl

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger (Eds.). 1106–1114. http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks

[19] Chung-Wei Lee, Wei Fang, Chih-Kuan Yeh, and Yu-Chiang Frank Wang. 2018. Multi-Label Zero-Shot Learning With Structured Knowledge Graphs. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 1576–1585. DOI: http://dx.doi.org/10.1109/CVPR.2018.00170

[20] Ron Levie, Federico Monti, Xavier Bresson, and Michael M. Bronstein. 2019. CayleyNets: Graph Convolutional Neural Networks With Complex Rational Spectral Filters. *IEEE Trans. Signal Processing* 67, 1 (2019), 97–109. DOI: http://dx.doi.org/10.1109/TSP.2018.2879624

[21] Qing Li, Xiaojiang Peng, Yu Qiao, and Qiang Peng. 2019. Learning Category Correlations for Multi-label Image Recognition with Graph Networks. *CoRR* abs/1909.13005 (2019). arXiv:1909.13005 http://arxiv.org/abs/1909.13005

[22] Yining Li, Chen Huang, Chen Change Loy, and Xiaoou Tang. 2016. Human Attribute Recognition by Deep Hierarchical Contexts. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI (Lecture Notes in Computer Science)*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.), Vol. 9910. Springer, 684–700. DOI: http://dx.doi.org/10.1007/978-3-319-46466-4_41

[23] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V (Lecture Notes in Computer Science)*, David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.), Vol. 8693. Springer, 740–755. DOI: http://dx.doi.org/10.1007/978-3-319-10602-1_48

[24] Lingqiao Liu, Peng Wang, Chunhua Shen, Lei Wang, Anton van den Hengel, Chao Wang, and Heng Tao Shen. 2017. Compositional Model Based Fisher Vector Coding for Image Classification. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 12 (2017), 2335–2348. DOI: http://dx.doi.org/10.1109/TPAMI.2017.2651061

[25] Mateusz Malinowski and Mario Fritz. 2014. A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger (Eds.). 1682–1690. http://papers.nips.cc/paper/5411-a-multi-world-approach-to-question-answering-about-real-world-scenes-based-on-uncertain-input

[26] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1301.3781

[27] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, Alessandro Moschitti, Bo Pang, and Walter Daelemans (Eds.). ACL, 1532–1543. DOI: http://dx.doi.org/10.3115/v1/d14-1162

[28] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1409.1556

[29] Jiaxiang Tang, Wei Hu, Xiang Gao, and Zongming Guo. 2019. Joint Learning of Graph Representation and Node Features in Graph Convolutional Neural Networks. *CoRR* abs/1909.04931 (2019). arXiv:1909.04931 http://arxiv.org/abs/1909.04931

[30] Konstantinos Trohidis, Grigorios Tsoumakas, George Kalliris, and Ioannis P. Vlahavas. 2008. Multi-Label Classification of Music into Emotions. In *ISMIR 2008, 9th International Conference on Music Information Retrieval, Drexel University, Philadelphia, PA, USA, September 14-18, 2008*, Juan Pablo Bello, Elaine Chew, and Douglas Turnbull (Eds.). 325–330. http://ismir2008.ismir.net/papers/ISMIR2008_275.pdf

[31] Grigorios Tsoumakas and Ioannis Katakis. 2007. Multi-Label Classification: An Overview. *IJDWM* 3, 3 (2007), 1–13. DOI: http://dx.doi.org/10.4018/jdwm.2007070101

[32] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. 2016. CNN-RNN: A Unified Framework for Multi-label Image Classification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2285–2294. DOI: http://dx.doi.org/10.1109/CVPR.2016.251

[33] Zhouxia Wang, Tianshui Chen, Guanbin Li, Ruijia Xu, and Liang Lin. 2017. Multi-label Image Recognition by Recurrently Discovering Attentional Regions. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 464–472. DOI: http://dx.doi.org/10.1109/ICCV.2017.58

[34] Yunchao Wei, Wei Xia, Min Lin, Junshi Huang, Bingbing Ni, Jian Dong, Yao Zhao, and Shuicheng Yan. 2016. HCP: A Flexible CNN Framework for Multi-Label Image Classification. *IEEE Trans. Pattern Anal. Mach. Intell.* 38, 9 (2016), 1901–1907. DOI: http://dx.doi.org/10.1109/TPAMI.2015.2491929

[35] Ruiqing Xu, Chao Li, Junchi Yan, Cheng Deng, and Xianglong Liu. 2019. Graph Convolutional Network Hashing for Cross-Modal Retrieval. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, Sarit Kraus (Ed.). ijcai.org, 982–988. DOI: http://dx.doi.org/10.24963/ijcai.2019/138

[36] Hao Yang, Joey Tianyi Zhou, Yu Zhang, Bin-Bin Gao, Jianxin Wu, and Jianfei Cai. 2016. Exploit Bounding Box Annotations for Multi-Label Object Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 280–288. DOI: http://dx.doi.org/10.1109/CVPR.2016.37

[37] Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Graph Convolutional Networks for Text Classification. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*. AAAI Press, 7370–7377. DOI: http://dx.doi.org/10.1609/aaai.v33i01.33017370

[38] Jing Yu, Yuhang Lu, Zengchang Qin, Weifeng Zhang, Yanbing Liu, Jianlong Tan, and Li Guo. 2018. Modeling Text with Graph Convolutional Network for Cross-Modal Information Retrieval. In *Advances in Multimedia Information Processing - PCM 2018 - 19th Pacific-Rim Conference on Multimedia, Hefei, China, September 21-22, 2018, Proceedings, Part I (Lecture Notes in Computer Science)*, Richang Hong, Wen-Huang Cheng, Toshihiko Yamasaki, Meng Wang, and Chong-Wah Ngo (Eds.), Vol. 11164. Springer, 223–234. DOI: http://dx.doi.org/10.1007/978-3-030-00776-8_21

[39] Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. 2018. Beyond Bilinear: Generalized Multimodal Factorized High-Order Pooling for Visual Question Answering. *IEEE Trans. Neural Netw. Learning Syst.* 29, 12 (2018), 5947–5959. DOI: http://dx.doi.org/10.1109/TNNLS.2018.2817340

[40] Min-Ling Zhang and Zhi-Hua Zhou. 2014. A Review on Multi-Label Learning Algorithms. *IEEE Trans. Knowl. Data Eng.* 26, 8 (2014), 1819–1837. DOI: http://dx.doi.org/10.1109/TKDE.2013.39

[41] Feng Zhu, Hongsheng Li, Wanli Ouyang, Nenghai Yu, and Xiaogang Wang. 2017. Learning Spatial Regularization with Image-Level Supervisions for Multi-label Image Classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2027–2036. DOI: http://dx.doi.org/10.1109/CVPR.2017.219