MULTI LABEL IMAGE CLASSIFICATION USING ADAPTIVE GRAPH CONVOLUTIONAL NETWORKS (ML-AGCN)

Inder Pal Singh, Enjie Ghorbel, Oyebade Oyedotun, Djamila Aouada

Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg, Luxembourg

ABSTRACT

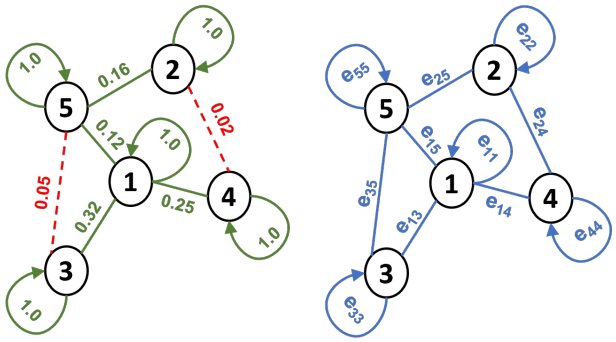
In this paper, a novel graph-based approach for multi-label image classification called Multi-Label Adaptive Graph Convolutional Network (ML-AGCN) is introduced. Graph-based methods have shown great potential in the field of multi-label classification. However, these approaches heuristically fix the graph topology for modeling label dependencies, which might be not optimal. To handle that, we propose to learn the topology in an end-to-end manner. Specifically, we incorporate an attention-based mechanism for estimating the pairwise importance between graph nodes and a similarity-based mechanism for conserving the feature similarity between different nodes. This offers a more flexible way for adaptively modeling the graph. Experimental results are reported on two well-known datasets, namely, MS-COCO and VG-500. Results show that ML-AGCN outperforms state-of-the-art methods while reducing the number of model parameters.

1. INTRODUCTION

Multi-label image classification can be defined as the task of predicting the set of object labels present in a given image. This topic has been widely studied by the computer vision community. This is mainly due to its practicality in numerous application areas including human attributes recognition [1], scene recognition [2] and multi-object recognition [3]. In fact, in comparison to single-label methods, multi-label image classification is more realistic since it assumes that a typical real-world image can comprise more than one object.

Given the tremendous advances in deep learning, most recent state-of-the-art methods rely on single-stream Convolutional Neural Networks (CNNs) [4, 5]. Despite their great performance, these approaches usually require a high number of layers to ensure effectiveness. As a result, the number of model parameters tends to increase, leading to a cumbersome architecture that is difficult to deploy in a memory-constrained environment.

Alternatively, a second class of methods exploits the prior knowledge related to the label correlations [6, 7, 8]. Indeed, some objects are more likely to appear together than others. As shown in [7], integrating such a strategy in a model can be a way for improving the scalability property. This means that a lower number of parameters would be needed for achieving comparable performance with one-stream models.



(a) Fixed graph $\tau = 0.1$ (b) Parameterized graph $\tau = 0$

Fig. 1. (a) An example of a fixed label graph with a threshold $\tau = 0.1$. Red edges indicate the ignored edges; (b) The proposed parameterized graph topology considering all edges.

Among the most popular multi-label image classification approaches modeling label correlations, one can mention graph-based methods [6, 7]. They are usually composed of the association of two subnets, namely, a traditional CNN that extracts discriminative features from an input image coupled with a Graph Convolutional Network (GCN) that generates N inter-dependent label classifiers, with N being the number of labels. GCN is an extension of CNN to graphs which has shown great performance in many tasks such as human pose estimation [9] and action recognition [10, 11]. The input graph is designed based on the label correlations, where each node corresponds to a label and each edge defines the co-occurrence probability between two labels. The generated classifiers are therefore used to predict the presence or not of the associated labels by considering the image features produced by the CNN subnet.

Despite their proven performance in terms of both precision and network size, these graph-based approaches remain impacted by three main limitations: (1) the topology of the graph is heuristically fixed. Specifically, it is com-

This work was funded by the Luxembourg National Research Fund (FNR), under the project reference BRIDGES2020/IS/14755859/MEET-A/Aouada.

puted based on the co-occurrence of labels in the training data which might be not the most adapted approach for the task of multi-label image classification; (2) a threshold is empirically set to ignore edges with low co-occurrence probability (see Fig. 1(a)). This means that rare co-occurrences are automatically considered to be noisy. Although this might be true in many cases, assuming that any rare event corresponds to noise does not always hold; and (3) it has been theoretically and empirically proved in [12] that the GCN aggregation procedure tends to destroy the node similarity in the original feature space, potentially leading to a decrease in terms of precision.

Herein, we posit that by integrating adequate mechanisms in graph-based approaches for addressing the aforementioned issues, it should be possible to reduce the size of the network, while achieving competitive performance.

In this paper, we propose an adaptive attention-based graph for multi-label image classification that we refer to as *Multi Label Adaptive Graph Convolutional Network (ML-AGCN)*. As described in Fig. 1(b), the idea consists in parametrizing two types of graphs in an end-to-end manner without the application of any threshold. On the one hand, the first one quantifies the connectivity importance of each node pair by learning an attention score similar to Graph Attention Network (GAT) [13], allowing more flexibility. On the other hand, the second graph is learned by considering the similarity between feature nodes; therefore avoiding an undesired loss of information through the convolutions. The proposed approach is tested on two well-known multi-label image datasets. The results suggest that our method is able to compete with the state-of-the-art while further reducing the number of parameters. The remaining of the paper is organized as follows: Section 2 provides the problem formulation. Section 3 introduces the proposed approach. Section 4 describes the experiments, and finally Section 5 concludes this work.

2. BACKGROUND: GRAPH-BASED APPROACHES FOR MULTI-LABEL IMAGE CLASSIFICATION

Given an input image I, the aim of multi-label image classification is to estimate a function f that predicts the presence or not of labels belonging to a set $\mathcal{L} = \{1, ..., N\}$. This can be written as follows,

$$\begin{aligned} f &: \mathbb{R}^{w \times h} \rightarrow [\![0,1]\!]^N \\ I &\mapsto \mathbf{y} = (y_i)_{i \in \mathcal{L}}, \end{aligned}$$

with w and h respectively the pixel-wise width and height of the image. Note that $y_i = 1$ if the label i is present in I, otherwise $y_i = 0$. As discussed in Section 1, graph-based multi-label methods such as ML-GCN [6] and IML-GCN [7] are composed of two branches. The first one is based on an out-off-shelf CNN model allowing the extraction of discriminative image representations. In particular, ML-GCN and IML-GCN incorporate respectively a ResNet-101 [4] and a TResNet-M [5]. The latter consists in an efficient version of ResNet-50. The second branch based on a standard GCN aims at generating N inter-dependant binary classifiers. Let us denote the input graph by $\mathcal{G} = \{V, E, \mathbf{F}\}$, with $V = [v_1, v_2, ..., v_N]$ the set of vertices such that v_i corresponds to the vertex associated to the label, $E = [e_1, e_2, ..., e_M]$ the set formed by M edges connecting the vertices and $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_N]$ the vertex features such that $\mathbf{f}_i \in \mathbb{R}^d$ represents the features of the vertex i. Let $\mathbf{A} \in \mathbb{R}^{N \times N}$ be the adjacency matrix defining the topology of the graph. $\mathbf{A}$ is computed by considering the co-occurrence probability of labels. Furthermore, a threshold τ that is empirically fixed is used to ignore rare co-occurrences that are considered as noisy. For $i, j \in \mathcal{L}$,

$$\mathbf{A}_{ij} = \begin{cases} 0, \text{ if } P_{ij} < \tau, \\ 1, \text{ if } P_{ij} \geq \tau \end{cases}, \tag{1}$$

where $P_{ij} = P(j|i)$ is the co-occurrence probability that the label j appears given that i is already present.

Then, assuming that $\mathbf{F}^l \in \mathbb{R}^{n \times d^l}$ encodes the input vertex features of the l^{th} layer, the GCN computes the node features of the $(l+1)^{th}$ layer $\mathbf{F}^{l+1} \in \mathbb{R}^{n \times d^{l+1}}$ as follows,

$$\mathbf{F}^{l+1} = h(\mathbf{A}\mathbf{F}^l\mathbf{W}^l), \tag{2}$$

with h a non-linear activation function mostly chosen as a Leaky Rectified Linear Unit (Leaky ReLU), $\mathbf{W}^l \in \mathbb{R}^{d^l \times d^{l+1}}$ the learned weight matrix of layer l. Note that $\mathbf{A}$ is normalized before applying Eq. (2). Finally, the vertex features produced by the last layer form the N inter-dependent classifiers.

In line with the limitations presented in Section 1, it can be observed that: (1) the adjacency matrix $\mathbf{A}$ incorporating the label correlation information is pre-computed independently from the training process; (2) a simple thresholding is applied for ignoring rare co-occurrences; (3) the successive aggregation of the neighboring nodes might lead to the loss of node similarity information present in the initial feature space, as highlighted in GCN [12].

3. MULTI-LABEL ADAPTIVE GRAPH CONVOLUTIONAL NETWORK

In order to overcome these issues, we propose a new graph-based multi-label approach that we call ML-AGCN.

3.1. Overview of the proposed approach

As in [6, 7], our network is composed of two main subnets: one CNN subnet that extracts discriminative features from an input image and a GCN-based network that allows learning N interdependent classifiers based on label correlations. As shown in Fig 2., similar to [7], we use a smaller version of TResNet [5] called TResNet-M as a CNN subnet. TResNet has been introduced to boost the neural network efficiency by fully exploiting the GPU capabilities. However, the

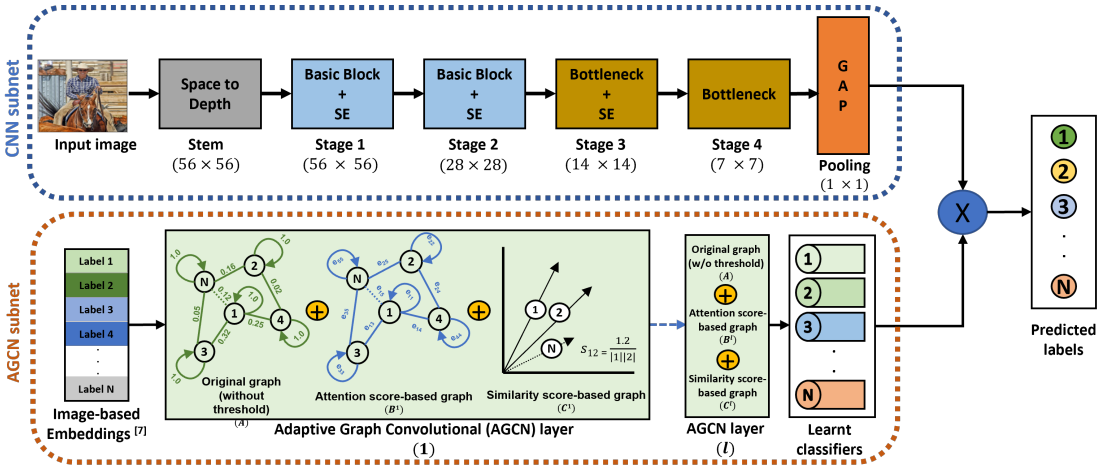


Fig. 2. Architecture of the proposed ML-AGCN. The CNN subnet extracts the discriminative features from an input image while the GCN subnet learns an adaptive graph based on the computed attention weights ($\mathbf{B}^1$) and similarity measures between the vertices ($\mathbf{C}^1$), then generates interdependent classifiers which are directly applied to the learned image representations.

graph-based subnet entitled Adaptive Graph Convolutional Network differs from traditional GCN employed in previous methods [7, 6]. Note that, as in [7], we make use of the Asymmetric Loss (ASL) [14] and employ the same image-based embeddings as node features. More details about this subnet are given in Section 3.2.

3.2. GCN-based subnet: Adaptive Graph Convolutional layer (AGCN)

To overcome the limitations presented in Section 2, we propose an Adaptive GCN. The idea consists in learning in an end-to manner the topology of the graph by redefining equation (2) as follows,

$$\mathbf{F}^{l+1} = h((\mathbf{A} + \mathbf{B}^{(l)} + \mathbf{C}^{(l)})\mathbf{F}^{l}\mathbf{W}^{l}). \tag{3}$$

This means that the structure of the graph depends on three different components, namely, the original adjacency matrix $\mathbf{A}$ defined as in [6][1], the l^{th} layer attention-based adjacency matrix $\mathbf{B}^{(l)}$ and the l^{th} layer similarity-based adjacency matrix $\mathbf{C}^{(l)}$. Note that $\mathbf{B}^{(l)}$ and $\mathbf{C}^{(l)}$ vary from a layer to another, while $\mathbf{A}$ is fixed. The computation of $\mathbf{B}^{(l)}$ and $\mathbf{C}^{(l)}$ is described below.

3.2.1. *Attention-based Adjacency Matrix*

Instead of discarding rare co-occurrences from the adjacency matrix, $\mathbf{B}^{(l)} = (b_{ij}^{(l)})_{i,j\in\mathcal{L}}$ aims at incorporating an attention mechanism which defines the importance of each edge. For that purpose, an attention score e_{ij} for each pair of vertices (v_i, v_j) is first computed as in [13], such that,

$$e_{ij} = \text{LeakyReLU}(\mathbf{a}^{(l)^T}(W\mathbf{f}_i^{(l)} || W\mathbf{f}_j^{(l)})), \tag{4}$$

where $\mathbf{W} \in \mathbb{R}^{d^{(l+1)}\times d^{(l)}}$ is a learnable weight matrix, $\mathbf{a}^{(l)^T} \in \mathbb{R}^{2d^{(l+1)}\times 1}$ are the learnable attention coefficients and $||$ refers to the concatenation operation. Then, a softmax function is applied to the attention scores as shown below,

$$\alpha_{ij}^{(l)} = \frac{\exp(e_{ij}^{(l)})}{\sum_{k\in\mathcal{N}(i)}\exp(e_{ik}^{(l)})}, \tag{5}$$

where $\mathcal{N}(i)$ defines the neighbourhood of the node i and α_{ij} is the normalized attention score. Finally, in order to preserve the self-importance of i during the aggregation, the attention-based adjacency matrix $\mathbf{B}^{(l)} = (b_{ij}^{(l)})_{i,j\in\mathcal{L}}$ can be computed by following operation,

$$\begin{cases} b_{ij}^{(l)} = \alpha_{ij}^{(l)} + \max\limits_{k\in\mathcal{L}}(\alpha_{ik}^{(l)}) \text{ if } i = j \\ b_{ij}^{(l)} = \alpha_{ij}^{(l)} \text{ if } i \neq j \end{cases}. \tag{6}$$

3.2.2. *Similarity-based Adjacency Matrix*

As mentioned in Section 1, since the aggregation procedure of GCN tends to destroy the node similarity [12], we propose to use a node-similarity preserving matrix $\mathbf{C}^{(l)} = (c_{ij}^{(l)})_{i,j\in\mathcal{L}}$ by computing a cosine similarity $c_{ij}^{(l)}$ for each pair of vertices (v_i, v_j) such that,

$$c_{ij}^{(l)} = \frac{\mathbf{f}_i^{(l)}.\mathbf{f}_j^{(l)}}{\|\mathbf{f}_i^{(l)}\|\|\mathbf{f}_j^{(l)}\|}, \tag{7}$$

where $\|.\|$ denotes the L_2 Euclidean norm.

[1] In this case, we do not apply any threshold for ignoring edges with low probabilities.

Table 1. Comparison with state-of-the-art methods on the MS-COCO dataset.

Method	#Parameters	mAP	CP	CR	CF1	OP	OR	OF1
CNN-RNN [18]	66.2 M	61.2	-	-	-	-	-	-
ResNet101 [4]	44.5M	77.3	80.2	66.7	72.8	83.9	70.8	76.8
Multi-Evidence [17]	∼47M	-	80.4	70.2	74.9	85.2	72.5	78.4
ML-GCN (2-layers) [6]	44.9M	83	85.1	72	78	85.8	75.4	80.3
ML-GCN (1-layer)*† [6]	43.1	80.9	82.9	69.7	75.8	84.8	73.6	78.8
SSGRL [19]	92.2M	83.8	**89.9**	68.5	76.8	**91.3**	70.8	79.7
KGGR [15]	∼45M	84.3	85.6	72.7	78.6	87.1	75.6	80.9
C-Tran [8]	120M	85.1	86.3	74.3	79.9	87.7	76.5	81.7
ASL (TResNetM) [14]	**29.5M**	81.8	82.1	72.6	76.4	83.1	76.1	79.4
ASL (TResNetL) [14]	53.8M	86.6	87.4	76.4	81.4	88.1	79.2	81.8
IML-GCN (2-layers) [7]	31.5M	86.6	78.8	**82.6**	80.2	79.0	**85.1**	81.9
IML-GCN (1-layer)*† [7]	29.5M	81.3	81.3	72.2	76.0	86.7	77.9	82.1
Ours - ML-AGCN (2-layers)	35.9M	**86.9**	86.2	78.3	**81.7**	87.2	80.7	**83.8**
Ours - ML-AGCN (1-layer)*	29.9M	86.6	79.6	82.4	80.7	79.8	84.5	82.1

*Graph-based approaches within 1-layer GCN setting

†Reproduced results

4. EXPERIMENTS

In this section, we report the obtained results on MS-COCO [3] and VG-500 [15] datasets. MS-COCO is a large-scale multi-label image dataset that provides 122,118 images with a total of 80 categories. VG-500 [16] consists of 108,077 images with 500 objects. Following the conventional evaluation protocol used for MS-COCO [17], we report the following: mean Average Precision (mAP), average per-Class Precision (CP) and Overall Precision (OP), average per-Class Recall (CR), and Overall Recall (OR), average per-Class F1-score (CF1) and Overall F1-score (OF). For VG-500, we report the standard mean Average Prevision (mAP). For both datasets, We also report the number of model parameters. In order to support our initial claim stating that it is possible to maintain the performance while reducing the number of layers and parameters, we propose two experimental settings for ML-GCN, IML-GCN and ML-AGCN. In the first setting, we keep the same depth of the GCN subnet used in graph-based approaches, i.e., 2 hidden layers. However, in the second setting, we reduce it to 1. Note that we do not reproduce the results for ML-GCN on VG-500 given the unavailability of label word embeddings.

4.1. Comparison with state-of-the-art

As shown in Table 1, the proposed approach outperforms state-of-the-art approaches including graph-based methods (ML-GCN and IML) on MS-COCO. Indeed, we achieve the best results in terms of precision with 86.9% mAP. The relevance of our approach is also confirmed in Table 2 reporting the results on VG-500. Achieving the second best results after C-Tran [8] with an mAP of 37.9% against 38.4%, it can be noted that our network is almost 4 times smaller. Also, the proposed ML-AGCN maintains almost the same performance when reducing the number of layers to 1 (-0.3% on MS-COCO and +0.8% on VG-500), in contrast to ML-GCN (-2.1% on MS-COCO) and IML-GCN (-5.3% on MS-COCO and -17.3% on VG-500).

Table 2. Comparisons with state-of-the-art methods on the VG-500 dataset.

Method	# Parameters	mAP (%)
ResNet-101 [4]	44.5M	30.9
ML-GCN [6]	44.9M	32.6
ASL (TResNetM)† [14]	**29.5M**	33.6
ASL (TResNetL)† [14]	54.8M	34.7
C-Tran [8]‡	120M‡	38.4‡
IML-GCN (2-layers) [7])	32.1M	34.5
IML-GCN (1-layer)†* [7])	30.6M	17.2†*
Ours - ML-AGCN (2-layers)	37.4M	**37.1**
Ours - ML-AGCN (1-layer)*	32.7M	**37.9***

‡The model is roughly 273% larger than our proposal

*Graph-based approaches within 1-layer GCN setting

†Reproduced results

Table 3. Ablation study: Impact of each learned graph on the performance on the MS-COCO and VG-500 dataset.

Input Graph for the GCN (1-layer)	MS-COCO (mAP)	VG-500 (mAP)
Original graph ($\mathbf{A}$)	81.1	17.2
+ Attention-based graph ($\mathbf{A}+\mathbf{B}$)	86.6 (**+5.1%**)	37.5(**+20.3%**)
+ Node-preserving graph ($\mathbf{A}+\mathbf{B}+\mathbf{C}$)	86.7 (**+0.1%**)	37.9 (**+0.4%**)

4.2. Ablation study

Table 3 reports the quantitative contribution of the proposed attention-based and similarity-based graphs computed within a 1-layer GCN subnet setup. It can be observed that a fixed topology-based graph, similar to IML-GCN [7] presents a lower mAP. However, with the inclusion of our proposed attention-based parameterized graph ($\mathbf{B}$), the performance improves by 5.1% and 20.3% for MS-COCO and VG-500, respectively. A slighter increase in the mAP can be seen after adding the node similarity-preserving graph ($\mathbf{C}$). This confirms the usefulness of the two matrices, especially the use of an attention-mechanism.

5. CONCLUSION

Integrating GCN with existing CNN-based approaches to exploit the prior knowledge of label correlations has been shown to be a good practice for tackling the multi-label image classification problem. However, the topology of the input graph for GCN is heuristically fixed and a threshold is empirically chosen to ignore edges corresponding to rare co-occurrences. Furthermore, through the theoretical and empirical analysis [12], it has been shown that the convolution process of the GCN might destroy the node similarities in the initial feature space. As such, this paper proposes an Adaptive Graph Convolutional Network that adaptively learns two types of graphs incorporating the connectivity importance and the node similarity. We show that the proposed approach achieves state-of-the-art results on MS-COCO. Furthermore, it is able to generate classifiers with competitive prediction scores with the use of only one layer, in contrast to previous graph-based approaches.

6. REFERENCES

[1] Yining Li, Chen Huang, Chen Change Loy, and Xiaoou Tang, "Human attribute recognition by deep hierarchical contexts," in *European conference on computer vision*. Springer, 2016.

[2] Jing Shao, Kai Kang, Chen Change Loy, and Xiaogang Wang, "Deeply learned attributes for crowded scene understanding," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[3] Sean Bell, C. Lawrence Zitnick, Kavita Bala, and Ross Girshick, "Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, June 2016.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[5] Tal Ridnik, Hussam Lawen, Asaf Noy, Emanuel Ben Baruch, Gilad Sharir, and Itamar Friedman, "Tresnet: High performance gpu-dedicated architecture," in *IEEE Winter Conference on Applications of Computer Vision*, January 2021.

[6] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo, "Multi-label image recognition with graph convolutional networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5177–5186.

[7] Inder Pal Singh, Oyebade Oyedotun, Enjie Ghorbel, and Djamila Aouada, "Iml-gcn: Improved multi-label graph convolutional network for efficient yet precise image classification," in *AAAI Workshop Program-Deep Learning on Graphs: Methods and Applications*, 2022.

[8] Jack Lanchantin, Tianlu Wang, Vicente Ordonez, and Yanjun Qi, "General multi-label image classification with transformers," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[9] Yanrui Bin, Zhao-Min Chen, Xiu-Shen Wei, Xinya Chen, Changxin Gao, and Nong Sang, "Structure-aware human pose estimation with graph convolutional networks," *Pattern Recognition*, vol. 106, pp. 107410, 2020.

[10] Konstantinos Papadopoulos, Enjie Ghorbel, Djamila Aouada, and Björn Ottersten, "Vertex feature encoding and hierarchical temporal modeling in a spatio-temporal graph convolutional network for action recognition," in *25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 452–458.

[11] Konstantinos Papadopoulos, Enjie Ghorbel, Oyebade Oyedotun, Djamila Aouada, and Björn Ottersten, "Deepvi: A novel framework for learning deep view-invariant human action representations using a single rgb camera," in *IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 2020.

[12] Wei Jin, Tyler Derr, Yiqi Wang, Yao Ma, Zitao Liu, and Jiliang Tang, "Node similarity preserving graph convolutional networks," in *ACM International Conference on Web Search and Data Mining*, 2021.

[13] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[14] Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor, "Asymmetric loss for multi-label classification," in *IEEE International Conference on Computer Vision*, 2021.

[15] Tianshui Chen, Liang Lin, Xiaolu Hui, Riquan Chen, and Hefeng Wu, "Knowledge-guided multi-label few-shot learning for general image recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[16] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David Shamma, Michael Bernstein, and Fei-Fei Li, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International Journal of Computer Vision*, vol. 123, 05 2017.

[17] Weifeng Ge, Sibei Yang, and Yizhou Yu, "Multi-evidence filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[18] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu, "Cnn-rnn: A unified framework for multi-label image classification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[19] Tianshui Chen, Muxin Xu, Xiaolu Hui, Hefeng Wu, and Liang Lin, "Learning semantic-specific graph representation for multi-label image recognition," in *IEEE International Conference on Computer Vision*, 2019.