A Novel Framework for Semi-supervised Multiple-label Image Classification using Multi-stage CNN and Visual Attention Mechanism

Joseph James S*, Lakshmi C
Department of Computational Intelligence, Faculty of Engineering and Technology, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamil Nadu, India

***Abstract*—To train deep neural networks effectively, a lot of labeled data is typically needed. However, real-time applications make it difficult and expensive to acquire high-quality labels for the data because it takes skill and knowledge to accurately annotate multiple label images. In order to enhance classification performance, it is also crucial to extract image features from all potential objects of various sizes as well as the relationships between labels of numerous label images. The current approaches fall short in their ability to map the label dependencies and effectively classify the labels. They also perform poor to label the unlabeled images when small amount of labeled images available for classification. In order to solve these issues, we suggest a new framework for semi-supervised multiple object label classification using multi-stage Convolutional neural networks with visual attention (MSCNN)and GCN for label co-occurrence embedding(LCE) (MSCNN-LCE-MIC), which combines GCN and attention mechanism to concurrently capture local and global label dependencies throughout the entire image classification process. Four main modules make up MSCNN-LCE-MIC: (1) improved multi-label propagation method for labeling largely available unlabeled image; (2) a feature extraction module using multi-stage CNN with visual attention mechanism that focuses on the connections between labels and target regions to extract accurate features from each input image; (3) a label co-existence learning that applies GCN to discover the associations between different items to create embeddings of label co-occurrence; and (4) an integrated multi-modal fusion module. Numerous tests on MS-COCO and PASCAL VOC2007 show that MSCNN-LCE-MIC significantly improves classification efficiency on mAP 84.3% and 95.8% respectively when compared to the most recent existing methods.**

Keywords*—Semi-supervised; visual attention; multi-label; image classification; label propagation

I. INTRODUCTION

Multiple labels image classification has lately sparked a lot of interest in areas such as human attribution recognition, music mood categorization, and multi-object recognition. Multi-label image categorization, as compared to single-label image recognition, turns into a useful and difficult job that necessitates a deeper comprehension of the image objects because images from real world application always encompass numerous objects that contain extensive semantic information. Numerous of noteworthy works have been suggested to investigate the semantic connections between different labels and accomplish successful classification of multiple label images. There are two major groups into which these strategies fit. In contrast to the former which focuses on learning the regional correlations between target labels and true labels, the latter uses a (GCN) [1] to acquire the overall label relationships among different objects. This multi-label image, as seen in Fig. 1, is typical in real life and primarily includes three objects: "person," "tennis ball," and "tennis rocket." Each of these objects is annotated based on the target regions that have been highlighted in the image rather than other areas of it.

Wang et al. [2] and others [3, 4, 5] coupled recurrent neural network (RNN) and convolution neural network (CNN) to mutually describe the image label significance and label relationships, but they failed to take into account of the geographic setting of images. The main limitation of these techniques is that they ignore the intricate topology structure that exists between objects. This encourages study into methods for identifying and examining the label dependency in different approaches.

To directly model label dependencies, some methods built around RNN [2] or stochastic graph models [6, 7] have been suggested. The first approach views the multiple label problem of classification as a systematic reasoning problem that might have scaling issues due to its high computational complexity. Another work uses attention mechanisms to tacitly model the label correlations [8]. The global associations between labels that must be derived from information drawn from visuals other than the one being studied are still being ignored. Instead, they consider correlations between observed areas of an image, also referred to as regional associations. To address this issue, Zhu et al. [9] propose the Spatial Regularization Network (SRN), which learns an attention map for each label by using image-level supervisions and focused on the associated image region for every label. To update the complete network, however, they employ a subpar multi-step training workflow. Despite the fact that these methods [10] analyze label relationships with mechanisms of attention, they only take into account a small amount of local association among various target regions that show in a single image and ignore association of the labels distribution on every one of the training images.

Fig. 1. Multiple label images.

Fig. 1 shows that if two things are semantically associated in the real world, they are more probable to appear together in an image compared to when they are not. The ML-GCN architecture developed by Chen et al. [11] records the global label correlations on all training examples by employing GCN to generate label relationship embedding using the label statistics. It has been suggested that A-GCN [12] represent the label correlations in reference to ML-GCN by developing a method for label network construction. However, these two approaches use the dot product (DP) to finish the fusion of label co-occurrence embeddings and image features, severely impeding model convergence and multiple label image categorization performance improvement. In this work, we suggest a new multi-staged CNN with a visual attention mechanism to extract features and produce better feature representations from the training images. This method captures the local and global label dependency and speeds up model convergence. We use a more effective label propagation technique to label the unlabeled images and multi modal factorized bilinear pooling (MFB) [13] as an element of fusion. Then, we suggest an innovative semi-supervised multiple label classification of images model with an attention mechanism in order to successfully integrate visual characteristics and label co-occurrence embeddings (referred to as MSCNN-LCE-MIC). MSCNN-LCE-MIC is composed of four fundamental modules: cross-modal fusing with MFB; learning by label co-occurrence embedding with GCN; and feature extraction with multi-staged CNN, attention mechanisms, and improved label propagation techniques.

In the first module, we use CNN inspired by VGG-16 to learn the features of the images by creating a feature map for each label with multi-stage CNN, a visual attention mechanism and an improved label propagation technique to obtain the labels for unlabelled images. All labels are transformed into word vectors in the second module before being used as inputs to GCN to produce label co-occurrence embedding. The MFB component is finally integrated into our task in the module of cross-modal fusion, where it helps us effectively combine label embeddings and image characteristics to allow a complete classification system. Widespread tests on MS-COCO [14] and PASCAL VOC2007 [15] show that MSCNN-LCE-MIC significantly improves convergence efficiency and outperforms existing methods in terms of classification outcomes. The following are the paper's main contributions:

1) By combining the multi-staged CNN with a visual attention process for effective extraction of features and GCN to simultaneously detect regional label relationships in an image and universal label relationships among different items over the data distribution, we provide an innovative complete capable of training multi-label image categorization framework, MSCNN-LCE-MIC.

2) To effectively extract features of each object from an image, which will improve the performance of the classification model, we suggest a multi-staged CNN with a visual attention mechanism.

3) We incorporate an improved label propagation technique to add labels to the training data's unlabeled images, and we also incorporate the fusion component MFB into MSCNN-LCE-MIC to effectively combine features of image and embeddings of label co-occurrence. As a result, model convergence is considerably accelerated and classification performance is improved.

4) The proposed method, MSCNN-LCE-MIC, consistently outperforms earlier competing approaches and is simple to apply end-to-end. We put our approach to the test using benchmark datasets for multiple label image identification.

The remaining portions of the piece of writing are structured as follows: In Section II, we will review recent works related to the topic. Section III will cover the suggested framework for semi-supervised multiple label image classification, along with a comprehensive explanation of each component. Section IV will discuss the experimental design, the datasets used, the architecture, and analysis of the results obtained. Section V, reviews the findings and concludes the whole work with the proposal to future research options.

II. Related Works

A. MultipleLabel Image Classification

Due to the emergence of large-scale datasets like Image Net [16], MSCOCO [14], and PASCAL VOC [15], as well as the quick advancement of deep convolutional networks [3, 17], the efficiency of image categorization has seen a quick improvement. Extending deep CNNs for categorization of images with multiple labels has received a lot of attention. Image categorization using a single label techniques have advanced significantly with the quick creation of CNN-based models. Experts employed binary methods to classify images that have multiple labels through developing a binary classification algorithm for each label. To recognize multi-label photos, [18] apply pre-built features of ImageNet. Chatfield et al.'s [19] model success is improved by using the target dataset to create task-specific image characteristics. A deep network that excels on particular items is VeryDeep [4]. These methods, however, deal every object in the image separately and ignore the connections between various elements. To jointly describe the label significance and relationships, Wang et al. suggest CNN-RNN [2].

The SRN created by Zhu et al. [9] places an emphasis upon the associated image area associated with every label and uses image level oversights to obtain an attention map for each label. They are unable to finish end-to-end training because they update the entire network using a subpar multi-step training workflow. The study [20] suggests employing

GCAM to record the label relationships between diverse image transforms. Graph convolution network (GCN) is used by MLGCN [11] and AGCN [12] to produce label co-occurrence embedding for multiple label picture categorizations. Unfortunately, they use dot product (DP) to finish the process of fusing label co-occurrence embeddings and image features, substantially impeding system convergence and more effectively classify images with multiple labels.

B. Learning with Structured Graph

By utilizing graph propagation and reasoning, it is possible to model the relationships between labels in a useful fashion as well. In order to improve image categorization, [21] used neural networks to analyze data with a graph layout and provide a paradigm using GNNs to discover more characteristic relationships. In order to investigate the relationships between various labels of multiple label zero-shot learning, [22] used information graphs. For the study of graph-structured data, [1] introduced a GCN method, which utilized layer wise propagation to encode both graph data and node attributes. According to the aforementioned GCN, Semantic embeddings and classification associations were merged by Wang et al. [23] to predict the effectiveness of the visual classification for each group. Moreover, [11] provided a framework built on GCN that modeled the relationships between labels for multi-label categorization using a predefined graph.

C. Cross-modal Fusion

Recent years have seen a rise in interest in the disciplines of visual question and answering (VQA) [24, 25] and multi-modal counterfeit news discovery [26], which attempts to successfully combine vectors from several modules. These fields combine visual and linguistic representation derived from open, sizable linguistic or visual databases. Nevertheless, current techniques are unable to produce expressive image-text features or to comprehend the complex connections between these characteristics. The majority of available solutions merely integrate cross-modal embeddings using linear models such element-wise addition. It has been suggested that MCB and MLB significantly lower the calculation costs associated with the VQA fusion procedure to address this issue. However, there are two significant drawbacks: MLB requires too much iteration before it converges, while MCB can only produce good results if it collects multi-dimensional feature vectors. Additionally, [13] proposed MFB, which successfully combines image features and text embeddings while also noticeably accelerating model convergence. MFB first transforms vectors of high dimension into vectors of low dimension before combining the matching position element of cross modal vectors. By incorporating MFB to be trained on the multiple label picture characteristics, F-GCN [27] successfully addresses the cross-modal fusion problem, but it disregards each image's attention process.

The problems in the existing methods are lack of efficient feature extraction methods to handle morphological similarity issues and label dependency that exists between objects of an image.

The proposed work presents an innovative semi-supervised framework called MSCNN-LCE-MIC that acquires label relationship and their interdependence with GCNs in order to enhance the accuracy of multiple object label image categorization. The aforementioned structure learning methods served as inspiration for this framework. Our MSCNN-LCE-MIC explicitly stacks numerous GCN layers to build universal models for exploring the feature representations, in contrast with earlier structure learning methods. The main difference between MSCNN-LCE-MIC and rival approaches is that the proposed work uses an end-to-end methodology during the training stage to concurrently record localized label correlations within an image along with universal label correlations across multiple items in the data distribution. The label encoding and structured representation of graphs techniques used in MSCNN-LCE-MIC are more efficient at addressing the over-smoothing and scaling problems that are brought on by the substantial depth of GCN-based architecture.

III. Proposed Work

The general structure of our suggested framework, multi-stage CNN with label co-occurrence embedding and attention mechanism (MSCNN-LCE-MIC) for semi-supervised classification of multiple label images, is given away in Fig. 2. This framework includes (1) enhanced label propagation to obtain labels for unlabelled images, (2) a multi-stage CNN with visual attention to mine the each input image features, (2) a module for learning label co-occurrence embedding that is built on the GCN, and (3) a module for fusing the two cross modal vectors described above successfully. We describe our plan's workflow in more detail below.

A. Improved Label Propagation

The label propagation method (LPA) is a well-known clustering technique due to its popularity and freedom from parameter dependence. Label propagation is a technique for propagating labels from labeled to unlabeled picture data based on the association between the two object classes [28][29]. The similarity weights in conventional label propagation systems, on the other hand, may not be appropriate for subsequent label propagation because they distribute labels after a certain data graph generation process. This strategy offers benefits, but it also has some disadvantages. Another drawback of LPA is the unpredictability with which nodes are clustered, which increases instability and the creation of big communities. When nearby nodes are chosen at random from a list of fixed constant hops, a collection of disconnected nodes results. We suggest an LPA-based solution to overcome the problem of random community allocation and build better clusters. These variations improve the quality of communities that are found by taking advantage of node attribute values and link strength.

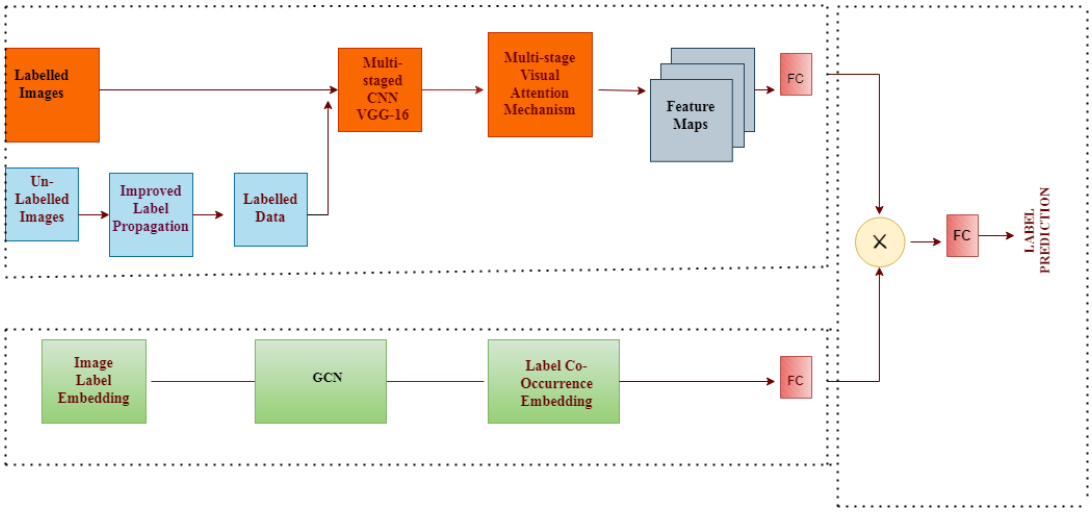


Fig. 2. Overall flow diagram of proposed method.

Using the feature data, a weighted graph is first created. The issue of random selection between nearest neighbors is then addressed via enhanced label propagation with dynamic connection strength metrics. The improved label propagation identifies the linked neighbors in the graph by using the heat kernel dispersion approach and its corresponding heat equation. The community arrangement of the node-attributed network is then returned by the program. In this form, the weighted common neighborhood (WCN) is employed as the link strength measure to award a community label (Murata &Moriyasu, 2007). The WCN value for each community label is then added together for the neighborhood. The system then chooses the community tag to apply to the selected node that has the highest total. The equation below describes how to measure the WCN connection strength.

$$Strength_{A,B} = \sum_{C \in \Gamma(A) \cap \Gamma(B)} w(A,C) + w(C,B) \quad (1)$$

The shared neighbors of nodes A and B are shown by the pair (A) (B). The edge weights that link the nodes A and B to their neighbor C are w (A, C) and w(C, B), respectively.

B. Multi-stage CNN

The projected model is based on an 18-layer CNN inspired by VGGNet [30]. However, as shown in Fig. 3, the proposed design classifies data by considering both high level and mid-level features. The second, third, and fourth network segments are used to extract these feature maps. The data from the first block is not directly used because it only contains minimal object label recognition filters. It is clear that the initial block's attributes made little of an impact. At this point, the networks can only understand and learn about simple textures, which is insufficient to define the essential traits required to differentiate objects.

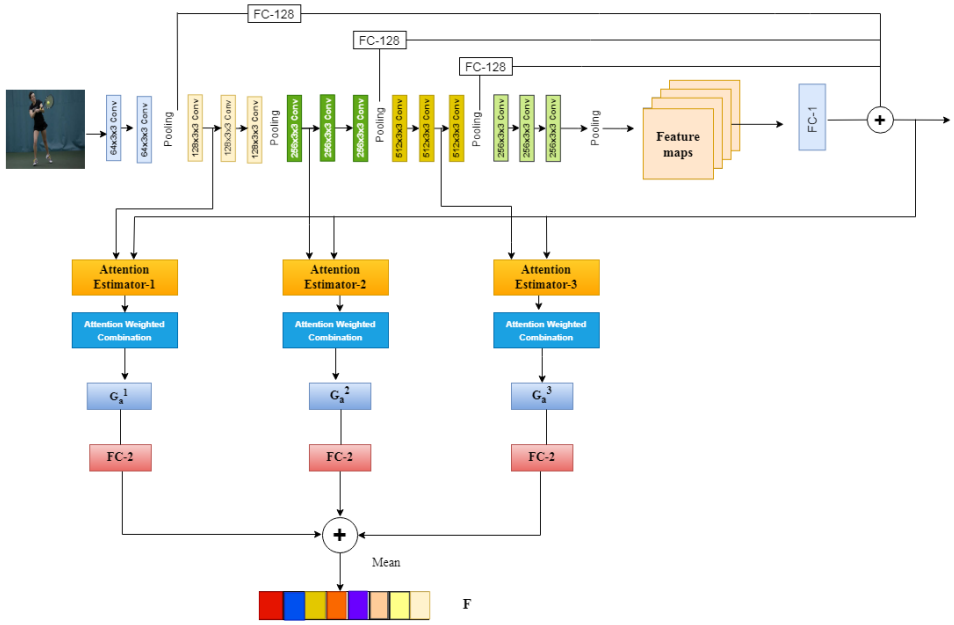


Fig. 3. Proposed multi-stage CNN with attention mechanism.

Furthermore, although mid-level layers in the second and third blocks produce local features, high-level layers in the fourth and fifth blocks greatly contribute to tiny object recognition via global characteristics. In this study, we propose MS-CNNs that combine local and global features in a predetermined way to produce the global feature vector g. Our networks, in contrast to conventional CNNs, generate a large number of essential characteristics that make the system robust to fluctuations in image quality and object occlusion problems. Also, a few pointless feature extraction filters are present in the same block, which is consistent with the discovery made regarding mid-level features in the prior section is given in eq.2.

$$(F_1^1, F_2^1, ...)^T \oplus (F_1^2, F_2^2, ...)^T = (F_1^1, F_2^1, ..., F_1^2, F_2^2, ...)^T \quad (2)$$

where $(F_1^1, F_2^1, ...)^T$ and $(F_1^2, F_2^2, ...)^T$ feature vectors that came from different network tiers. If the input is the result of Convolutional layers, vectorization is required before employing this operator.

C. Attention Mechanism

Attention estimators have been inserted after layers 7, 10, and 13. The attention estimator receives the layer 7 output and generates a "attention mask" consisting of integers between 0 and 1, then multiplies by output of the layer 7 to produce "g_a1". Following layers 10 and 13, the attention estimators go through a similar process, producing, respectively, g_a2 and g_a3. There was often another fully connected layer (FC) after the FC layer with the number "16," but it has been eliminated, leaving only the fully connected layer that comes after the dense layer at the network end to make the label classification. Instead, the three attention estimators' inputs are now handled by a new fully connected layer. The process of visual attention method for feature extraction is described as follows,

Step 1: The compatibility score C is calculated with use of the regional feature vector l and the universal feature vector g. The compatibility score is meant to be high when the local characteristics-defined image patch contains components of the dominating picture category. For instance, if the image contains a multiple objects, we assume that the global feature vector g adequately describes all possible object features. The patch that most closely resembles a particular object is also expected to produce local traits l that, when paired with g, will result in a high compatibility score.

$$C_i^s = \left\langle l_i^s, g \right\rangle, i\varepsilon\{1....n\} \quad (3)$$

The local feature vector l and the global feature vector g are simply added together. Be aware that while the local features l will change based on the convolutional layer (layers 7, 10, and 13), the global feature vector g will remain constant. Projecting g into l's lower-dimensional space will make sense if l and g are not the same dimensions.

Step 2: Calculate the Attention Weights a_i from the Compatibility Scores C shown in eq(4). The outcome is referred to as “a”, after that the compatibility scores C are compressed into the range of (0, 1) using a softmax.

$$a_i^s = \frac{\exp(C_i^s)}{\sum_j^n \exp(C_i^s)}, i\varepsilon\{1....n\} \quad (4)$$

Step 3: Determine Each Layer's Final Attention Mechanism Output using equation 5.

$$g_a^s = \sum_{i=1}^n a_i^s . l_i^s \quad (5)$$

Here, we calculate the attention mechanism's (g_a) final output for a certain layer using a weighted combination of the l for that particular layer (s). The attention weights "a" that we recently found are the weights that we employ.

Step 4: Create a categorization forecast based on final outcome of the attention module. We now want to select a classification using the attention outputs g_a that we just gathered for layers 7, 10, and 13. To acquire intermediate predictions, feed each attention output into a separate, completely connected layer. The final projections are then calculated by averaging these intermediate guesses.

D. Label Dependency Learning with GCN

According to [1] and [11], the label co-existence learning (LCL) module uses GCN of two-layers, with each layer taking the output of the layer before it and producing a new graph representation. Fig.4 shows the LCL module with GCN. The semantic encoding vectors $X = \{X_i\}_{i=1}^C$ and the equivalent association graph G are fed into the LCL component for the initial GCN layer. The graph representation and node feature may be easily combined by the LCL module in the convolution. The LCL module, which serves as the central part of the proposed MSCNN-LCE-MIC, intends to be trained on a set of classifier score $W\epsilon R^{c\,x\,d}$ to recalibrate early values for every label received from the image feature embedding module. In order to satisfy the requirements of the broadcast method, it is important to address two crucial challenges, namely word embedding and graph representation (5).

1) Word embedding: To retrieve the word embeddings for each of the labels across these multiple label image datasets, we use the 300-dim GloVe [31] algorithm learned on the Wikipedia dataset. We demonstrate that it is evenly efficient when using the suggested LCL method as demonstrated in Sections 4(c).To acquire the global label interdependence on the training set, we build a GCN with two layers. We use GloVe [31][33] to create 300 dimensional vectors for each of the objects in order to create the label embeddings matrix $Z\epsilon$ RCx300, following MLCGN [32].

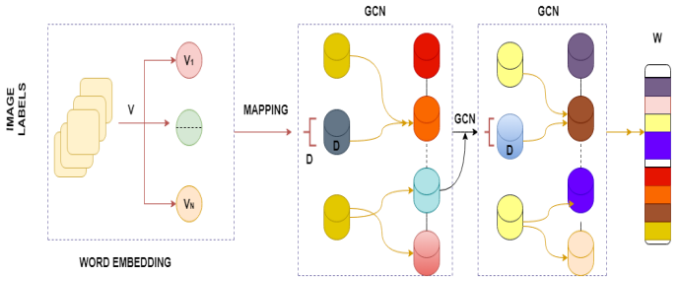


Fig. 4. Label Co-occurrence embedding network with GCN.

2) Graph representation: It goes without saying that the correlation matrix A affects how the node representations propagate. Instead of starting from scratch, researchers have suggested a number of methods for building the predetermined correlation matrix (explicit graph representations). For instance, WordNet [34] was used by Lee et al. [22] to develop their structured knowledge network. String matching was employed by Gao et al. [35] to plot concepts to nodes available in concept Net [36]. To express relationships between labels, these methods rely solely on semantic embedding and do not explicitly include information about interdependencies. Instead, we concentrate on the label co-existence matrix based on the training images, which is then used to represent the organized network composed of label associations by combining the label relationship data from neighboring nodes into a singular association matrix.

The following is the production procedure:

1) The first component of the label co-existence matrix P for the training image, we count the times at which pairs labels (L_i, L_j) first emerge.

$$P = \{p_i\}_{i=1,j=0}^{c} \in \Re^{cxc} \tag{6}$$

2) The initial label correlation matrix P' can be constructed using the matrix $R^{c\,x\,c}$.

$$P^{'} = \{p_i / N_i\}_{i=1,j=0}^{c} \in \Re^{cxc} \tag{7}$$

where N_i is the object label's i^{th} number. As a result, the graph constructed from such an asymmetric matrix P' is directed.

$$M_{i,j} = \begin{cases} & P_{i,j}^{'} < \varphi \\ 0, & P_{i,j}^{'} \geq \varphi \\ \lambda * \dfrac{P_{i,j}^{'}}{\sum_{j=0}^{c} P_j^{'} + \theta}, & \end{cases} \tag{8}$$

In order to safeguard label partnerships' specifics as stated in eq. (8), this work presents a nonlinear method for preparing matrix P' which can reduce noise (7), where φ is the noise filtration cutoff, 1^{e-6}, and M is the label association matrix employed in every GCN layer.

E. Multi-Label Classification Loss

Both the MS-COCO and the PASCAL VOC-2007 datasets exhibit the issue of class imbalance, which usually manifests as an inequity in the amount of negative and positive data. The suggested weighted cross entropy loss is taken into account as the multiple label categorization loss employed in our MSCNN-LCE-MIC to tackle this issue, and is described as:

$$Loss(b_i, l_i) = -wt_p \sum_{l_i=1} \log(\sigma(b_i)) - wt_n \sum_{l_i=0} \log(1-\sigma(b_i)) \tag{9}$$

Where, $wt_p = \dfrac{|po|+|Ne|+|1|}{|Po|+|1|}$ and $wt_n = \dfrac{|po|+|Ne|+|1|}{|Ne|+|1|}$, |Po| and |Ne| are The total amount of both positive and negative image labels in a batch, b_i is belief of each class label, σ is the sigmoid function.

Formally, for all i ϵ [1,C], we will fuse F and W_i to obtain Y^i, or the i^{th} component of the predicted label Y ϵ R^C, where W_i indicates the i^{th} row vector of W. This is done provided the I^{th} image feature vector F. First, using two f_C layers, W_i and F will be transformed into the corresponding m dimensional vectors M1 and M2. Additionally, M1 and M2 are cross-modal vectors that will be multiplied element-wise into an m-dimensional vector M1 M2 to enhance the interaction between these two embeddings. We further convert M1 ⊗ M2 into a m / g dimensional vector M via group sum-pooling to decrease parameter inflation and over-fitting, the elements in each group are represented by the letter g, to hasten convergence. The i^{th} constituent of Y is then obtained by creating a f_C layer. As a result, we are able to produce the full anticipated labels Y following fusion of C number of times. In order to create an end-to-end classification model, we employ the multiple label loss function by updating the Loss given in eq. (9) between predicted labels Y and the actual labels $Y^{'}$ ϵ {0, 1 }^{C} of I^{th} image.

IV. EXPERIMENTAL SETUP

A. Specifications of Implementation

PyTorch is used to perform all experiments. Each input image is resized to 448x448 before passing it to feature extraction component. Using the GloVe [37] model, each object is transformed to become a 300-dimensional vector that includes words in the label association embedding module. We set g = 2 to perform the group sum-pooling procedure and m = 358 to perform the fusion of the vectors of features and label co-existence embedded data with reference to the FGCN. A batch size of 32 is used when updating our network during the training procedure. The proposed model uses stochastic gradient descent (SGD) with 0.9 momentum, 10^{-4} weight decay and initial learning rate of 0.001.

1) Datasets: We conducted in-depth tests to confirm the effectiveness of MSCNN-LCE-MIC on MS-COCO [14] and PASCAL-VOC2007 [15]. To separate the datasets, we use the similar settings of MLGCN and FGCN. For further information, consult the references [5, 24].

2) Metrics for evaluation: We employ the following assessment metrics in accordance with mainstream techniques [9, 11]: (P-C) precision per class, (R-C) recall per-class, (mAP) mean average precision, F1 per class (F1-C) and F1 overall (F1-O). For fair comparisons, we also catalog the investigational findings on the top-3 classes of the categorization scores.

B. Investigational Results and Discussions

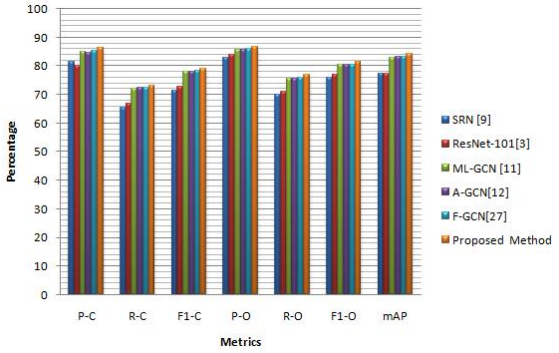
The convergence effectiveness and classification outcomes of MSCNN-LCE-MIC are compared to those of cutting-edge image classification methods.

1) Convergence efficiency: We track how the number of training epochs changed the mAP on the examination set in this section. We execute this experiment using the same parameters as ML-GCN, including SGD, batch size, learning rate, datasets, loss function, etc., to allow for fair comparisons. MSCNN-LCE-MIC has, on MS-COCO and PASCAL VOC2007, converged at the 25th and 23th epochs, respectively, and it produces superior mAP of 84.3% and 95.8%. Instead, MLGCN has not yet converged, and its mAP values are 45.8% and 75.7% lower than MSCNN-LCE-MIC at this time. Moreover, it will take the MLGCN roughly 200 epochs (more than 10 times as long as MSCNN-LCE-MIC) to complete its learning process. These findings show that multi-stage CNN, visual attention and cross module fusion significantly quickens model convergence and enhances performance of classification model.

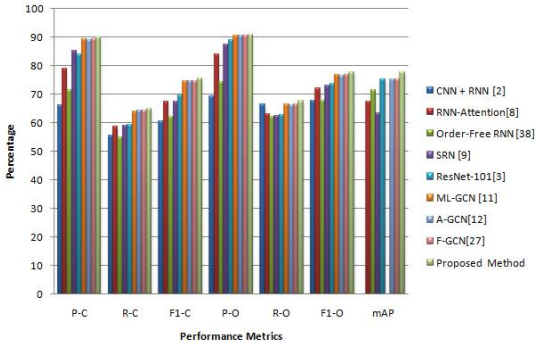
2) Results comparisons with the existing models

a) Evaluation results on MS-COCO: The most recent existing techniques are compared to MSCNN-LCE-MIC, including FGCN [27], AGCN [12], MLGCN [11], ResNet-101 [3], SRN [9], Order Free RNN [38], RNN with Attention [8], and CNN+RNN [2]. Fig. 5(a) and 5(b) shows the graphical representations of results obtained on MS-COCO dataset using proposed method. MSCNN-LCE-MIC nearly outperforms other options on all metrics which are shown in Table I and Table II. MSCNN-LCE-MIC significantly enhances the classification outcomes by 7% mAP in comparison to ResNet-101 baseline [3]. This occurrence shows that understanding the label dependency to produce more precise label features is greatly influenced by GCN. Also, as compared to those DP-based approaches, cross fusion module effectively completes the modal fusion to improve the efficiency of MSCNN-LCE-MIC. Additionally, based on the discrepancy between the earlier FGCN [27] and MLGCN [11], the multi-stage CNN with attention mechanism does in fact assist in extracting more precise image features, improving the effectiveness of categorization outcomes by 1.3% mAP.

b) Evaluation results on VOC2007: The state-of-the-art approaches are compared to MSCNN-LCE-MIC, including CNN+RNN [2], MLGCN [5], AGCN [6], ResNet-101 [8], Attention Reinforce [16], RNN with Attention [14], Very Deep [13], and FGCN [24]. The AP values and mAP values for each category are listed in Table III. With the exception of somewhat poorer performance on "bike", “bird” and "person" MSCNN-LCE-MIC outperforms other contenders overall in every category. The main cause is that the majority of earlier techniques ignored the local and global label dependencies, allowing them to solely focus on one or a few objects while ignoring the distribution of global labels. Our MSCNN-LCE-MIC, in contrast to existing approaches, considers both the regional label relationships in addition to the universal label relationships among diverse objects (multi-stage CNN with attention mechanism) within an image. MSCNN-LCE-MIC outperforms other methods for the remaining 17 objects. Although it appears that there is a trade-off between other approaches and our MSCNN-LCE-MIC in this phenomenon, we think that this global viewpoint is essential for multi-label picture categorization. The MSCNN-LCE-MIC produces a superior mAP value of 1.8% when compared to the existing method FGCN [27], which shows that both the multi-stage CNN with attention mechanism and the cross fusion module are involved and help to provide more accurate classification results. In Table IV of Section IV C, we also discuss how these modules have an impact.



(a)



(b)

Fig. 5. Performance comparison of proposed method on MS-COCO dataset (a) Overall labels (b) top-3 class.

TABLE I. ACCURACY(%) OBTAINED ON MS-COCO IMAGE DATASET

Method	P-C	R-C	F1-C	P-O	R-O	F1-O	mAP
[9]	81.6	65.4	71.2	82.7	69.9	75.8	77.1
[3]	80.2	66.7	72.8	83.9	70.8	76.8	77.3
[11]	85.1	72.0	78.0	85.8	75.4	80.3	83.0
[12]	84.7	72.3	78.0	85.6	75.5	80.3	83.1
[27]	85.4	72.4	78.3	86.0	75.7	80.5	83.2
MSCNN-LCE-MIC	86.3	73.0	79.1	86.8	76.9	81.5	84.3

TABLE II. TOP-3 ACCURACY (%) ON MS-COCO DATASET

Method	P-C	R-C	F1-C	P-O	R-O	F1-O	mAP
[2]	66.0	55.6	60.4	69.2	66.4	67.8	–
[8]	79.1	58.7	67.4	84.0	63.0	72.0	67.4
[38]	71.6	54.8	62.1	74.2	62.2	67.7	71.5
[9]	85.2	58.8	67.4	87.4	62.5	72.9	63.2
[3]	84.1	59.4	69.7	89.1	62.8	73.6	75.2
[11]	89.2	64.1	74.6	90.5	66.5	76.7	–
[12]	89.0	64.2	74.6	90.5	66.3	76.6	75.2
[27]	89.3	64.3	74.7	90.5	66.6	76.7	75.2
MSCNN-LCE-MIC	89.9	65.0	75.4	91.0	67.7	77.6	77.8

TABLE III. CLASSIFICATION ACCURACY RESULTS (%) ON VOC 2007 DATASET

Method	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mAP
CNN+RNN **[2]**	96.7	83.1	94.1	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
Very Deep [**4**]	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 **[3]**	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
RNN+Attention**[8]**	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
Attention-reinforce[39]	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
MLGCN [11]	99.5	98.5	96.8	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
AGCN[12]	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
FGCN[27]	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	98.9	93.7	94.0
MSCNN-LCE-MIC	**99.7**	**98.5**	**98.8**	**98.3**	**84.7**	**96.6**	**97.6**	**98.8**	**83.9**	**96.5**	**87.5**	**98.6**	**98.9**	**97.4**	**99.1**	**85.7**	**97.1**	**85.1**	**99.2**	**94.6**	**95.8**

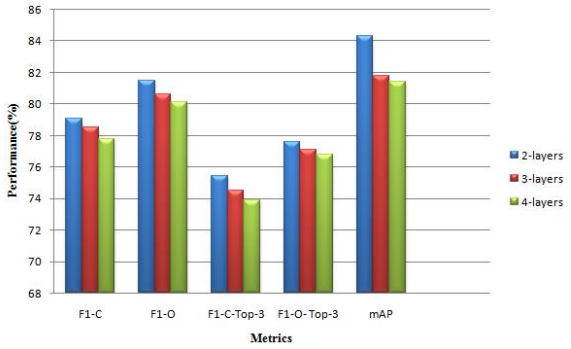
C. AblationStudies

We carry out ablation study to examine how important elements and parameter settings affect MSCNN-LCE-MIC.

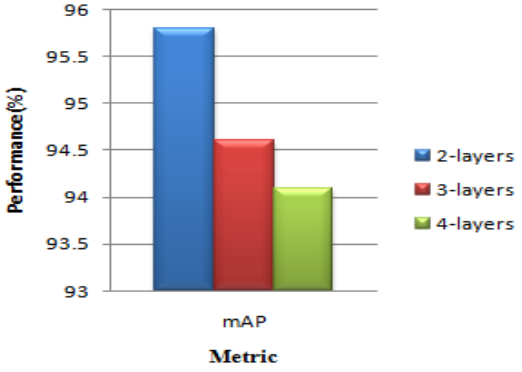
1) Classification performance of MSCNN-LCE-MIC with/without different modules: In this section, we examine how five crucial modules—namely, Multi-stage CNN, the visual attention method, the GCN, label propagation and the cross modal fusion affect our proposed model. Cross-modal fusion won't work if our suggested paradigm is used without GCN. Table IV displays the mAP results for MS-COCO and PASCAL VOC2007 with different module combinations. As can be seen, MSCNN-LCE-MIC performs best when all four of these modules are used at once. The evaluation findings show that any one module can enhance our model's mAP output. By incorporating the label dependencies between objects in global level, GCN especially improves MS-COCO and PASCAL VOC2007 mAP by 5.9% and 4.1%, respectively. In addition, the visual attention method boosts the mAP value by 0.3% on these two datasets while taking into account the label dependencies within an image. Finally, MFB keeps improving the classification outcomes with respective mAP improvements of 0.4% and 0.5%. These outcomes attest to the potency of our strategy, which notably benefits from both local and global level label dependencies, also the cross-modal fusion to boost the accuracy of classification.

2) GCN with different layers: The change in performance is shown in Table V and Fig. 6 after designing two (1024-2048), three (1024-1024-2048), and four (1024-1024-1024-2048) GCN layers in this section. With MS-COCO, MSCNN-LCE-MIC achieves its best performance with two GCN layers (Fig. 6(a)), after which its performance on mAP would degrade as GCN layers are added. Similar to Fig. 6(b), MSCNN-LCE-MIC produced the best outcome (mAP) on PASCAL VOC2007 using a 2-layer GCN. The main reason is because adding more layers of GCN would have a major negative impact on the model's functionality by making it so that the output features of nodes are no longer recognizable during the propagation process. We therefore employ two GCN layers in order to acquire the label-occurrence embeddings.

3) Number of units g in group sum pooling: By using group sum-pooling, we reduce each **m** dimensional vector to a manageably small m / g dimensional vector. Changing g from 1 to 64 allows us to track how performance changes and the outcomes are exposed in Table VI. As can be shown, g = 2 improves the outcome on MS-COCO even though the improvement in mAP on PASCAL VOC2007 is less pronounced. We think that g = 2 is more appropriate for minimizing the dimension and convey the semantic significance of the top phrase, despite the fact that various values of g have a comparable effect.



(a)



(b)

Fig. 6. Change in performance of proposed method with different GCN layers.

4) Performance of improved label propagation: The PASCAL VOC-2007 is a sizable multiple label standard dataset gathered for a number of computer vision tasks, including captioning, recognition and segmentation. The test collection contains 77,980 images, 24640 number of objects. There are about 2.4 object labels per picture, and there are 20 different classes in which the objects are divided. The 5K samples are used for testing, while the 20K samples are divided into 5 batches and used as training data. From each batch, we took 4k labeled samples at random and used the remaining 16K examples as unlabeled data. For each batch, this process is repeated five times. The end accuracy is the average of these trials. 32 mini-batch sizes were employed. Table VII and Table VIII display the outcomes of enhanced label propagation using various epochs. The proposed label propagation method achieves 3% less error than the existing label propagation method.

TABLE IV. CLASSIFICATION PERFORMANCE WITH / WITHOUT DIFFERENT MODULES

Improved label propagation	Multi-stage CNN	Visual Attention	GCN	MFB	mAP	
					MS-COCO	VOC-2007
Yes	Yes	Yes	Yes	Yes	**84.3**	**95.8**
Yes	Yes	Yes	No	No	83.5	94.4
Yes	No	No	No	No	77.4	90.6
No	No	Yes	Yes	Yes	83.7	94.6
No	Yes	No	Yes	Yes	84.0	94.6
No	Yes	Yes	No	No	77.2	90.2

TABLE V. PERFORMANCE COMPARISON ON GCN WITH VARIOUS LAYERS

Number of layers	MS-COCO					VOC-2007
	F1-C	F1-O	F1-C-top-3	F1-O-top-3	mAP	mAP
2-layers	79.1	81.5	75.4	77.6	84.3	95.8
3-layers	78.5	80.6	74.5	77.1	81.8	94.6
4-layers	77.8	80.1	73.9	76.8	81.4	94.1

TABLE VI. MODEL PERFORMANCE WITH DIFFERENT G VALUE

Dimensions of g	mAP	
	MS-COCO	VOC-2007
1	83.8	95.6
2	84.3	95.8
4	84.1	95.6
8	84.0	95.4
16	83.7	95.2
32	83.5	95.1
64	83.4	94.9

TABLE VII. ERROR RATE COMPARISON ON CIFAR-10

Dataset	CIFAR-10			
Labeled images	1000	2000	3000	4000
LPA [40]	22.02±0.88	15.66±0.35	-	12.69±0.29
Improved LA(ours)	18.24±0.50	11.78±0.65	10.80±0.30	9.40±0.50

TABLE VIII. ERROR RATE COMPARISON ON PASCAL VOC-2007

Dataset	PASCAL VOC-2007		
Labeled images	2000	4000	6000
LPA[40]	36.72±0.45	27.64±0.55	20.46±75
Improved LPA (ours)	30.33±0.25	20.25±0.56	17.70±0.44

V. CONCLUSION AND FUTURE WORK

This paper argues that most of the images in large scale datasets are unlabeled; the labels have dependency and morphological similarity issues. The existing methods using traditional convolution technique may fail to extract features efficiently and classify labels accurately due to large unlabeled data. We construct MSCNN-LCE-MIC, which combines multi-stage CNN with visual attention and GCN to concurrently collect global and regional level dependencies. A feature extraction component with multi-stage CNN and visual attention technique helps to create the most precise features of each image by concentrating on the relationships among labels and regions of target which solves the problem of morphological similarity issues exists between objects. MSCNN-LCE-MIC primarily consists of four key modules: improved label propagation technique to find labels of large volumes of unlabeled images available in real time applications; learning tool for label co-occurrence with GCN; and multi-stage CNN with attention mechanism. Comprehensive tests on MSCOCO and VOC2007 show that MSCNN-LCE-MIC significantly improves the effectiveness of our suggested framework and yields superior classification outcomes than the existing methods in the field. Experimental results demonstrate that the MSCNN-LCE-MIC method achieves higher mAP of 3.6% on MS-COCO dataset and 1.8% mAP on PASCALVOC-2007 dataset.

In the future, instead of using label propagation to train a neural network model, a generative adversarial network model could be used to get more labeled images from unlabeled image data and also reduce the time complexity due fusion of multiple components.

REFERENCES

[1] Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. In: 5th international conference on learning representations, Toulon, France, April 24-26, 2017.

[2] Wang J, Yang Y, Mao J, Huang Z, Huang C, Xu W. CNN-RNN: A unified framework for multi-label image classification. In: 2016 IEEE conference on computer vision and pattern recognition, Las Vegas, NV, USA, June 27-30, 2016. p. 2285–94.

[3] Kaiming He, Xiangyu Zhang, ShaoqingRen, and Jian Sun. Deep residual learning for image recognition. In CVPR, pages 770–778, 2016.

[4] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, pages 1–8, 2015.

[5] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and ZbigniewWojna. Rethinking the inception architecture for computer vision. In CVPR, pages 2818–2826, 2016.

[6] Qiang Li, MaoyingQiao, Wei Bian, and Dacheng Tao. Conditional graphical lasso for multi-label image classification. In CVPR, pages 2977–2986, 2016.

[7] Xin Li, Feipeng Zhao, and YuhongGuo. Multi-label image classification with a probabilistic label enhancement model. In UAI, pages 1–10, 2014.

[8] Zhouxia Wang, Tianshui Chen, Guanbin Li, RuijiaXu, and Liang Lin. Multi-label image recognition by recurrently discovering attentional regions. In ICCV, pages 464–472, 2017.

[9] Zhu F, Li H, Ouyang W, Yu N, Wang X. Learning spatial regularization with image-level supervisions for multi-label image classification. In: 2017 IEEE conference on computer vision and pattern recognition, Honolulu, HI, USA, July 21-26, 2017. p. 2027–36.

[10] Mnih V, Heess N, Graves A, Kavukcuoglu K. Recurrent models of visual attention. In: Advances in neural information processing systems 27: annual conference on neural information processing systems, Montreal, Quebec, Canada, December 8-13, 2014. p. 2204–12.

[11] Chen Z, Wei X, Wang P, Guo Y. Multi-label image recognition with graph convolutional networks. In: IEEE conference on computer vision and pattern recognition, Long Beach, CA, USA, June 16-20, 2019. p. 5177–86.

[12] Li Q, Peng X, Qiao Y, Peng Q. Learning category correlations for multi-label image recognition with graph networks. 2019, CoRR abs/1909.13005.

[13] Yu Z, Yu J, Xiang C, Fan J, Tao D. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. IEEE Trans Neural Netw Learn Syst 2018;29(12):5947–59.

[14] Lin T, Maire M, Belongie SJ, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. Microsoft COCO: Common objects in context. In: Computer vision -ECCV 2014 - 13th European conference, Zurich, Switzerland, September 6-12, 2014. p. 740–55.

[15] Everingham M, Gool LV, Williams CKI, Winn JM, Zisserman A. The pascal visual object classes (VOC) challenge. Int J Comput Vis 2010;88(2):303–38.

[16] JiaDeng,Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In CVPR, pages 248–255, 2009.

[17] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In CVPR, pages 7132–7141, 2018.

[18] Razavian AS, Azizpour H, Sullivan J, Carlsson S. CNN features off-the-shelf: An astounding baseline for recognition. In: IEEE conference on computer vision and pattern recognition, Columbus, OH, USA, June 23-28, 2014. p. 512–9.

[19] Chatfield K, Simonyan K, Vedaldi A, Zisserman A. Return of the devil in the details: Delving deep into convolutional nets. In: British machine vision conference, Nottingham, UK, September 1-5, 2014.

[20] Wang Y, Xie Y, Liu Y, Fan L. G-CAM: Graph convolution network based class activation mapping for multi-label image recognition. In: ICMR '21: International conference on multimedia retrieval, Taipei, Taiwan, August 21-24, 2021. p. 322–30.

[21] K. Marino, R. Salakhutdinov, and A. Gupta, "The more you know: Using knowledge graphs for image classification," in Proc. Conf. Comput. Vision Pattern Recognit., 2016, pp. 20–28.

[22] C.-W. Lee,W. Fang, C.-K. Yeh, and Y.-C. FrankWang, "Multi-label zeroshot learning with structured knowledge graphs," in Proc. IEEE Conf. Comput. Vision Pattern Recognit., 2018, pp. 1576–1585.

[23] X. Wang, Y. Ye, and A. Gupta, "Zero-shot recognition via semantic embeddings and knowledge graphs," in Proc. IEEE Conf. Comput. Vision Pattern Recognit., 2018, pp. 6857–6866.

[24] Malinowski M, Fritz M. 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, Montreal, Quebec, Canada, December 8-13, 2014. p. 1682–90.

[25] Chen J, Zhang S, Zeng J, Zou F, Li Y-F, Liu T, Lu P. Multi-level, multi-modal interactions for visual question answering over text in images. World Wide Web 2021;1–17.

[26] Zeng J, Zhang Y, Ma X. Fake news detection for epidemic emergencies via deep correlations between text and images. Sustainable Cities and Society 2021;66:102652.

[27] Wang Y, Xie Y, Liu Y, Zhou K, Li X. Fast graph convolution network based multi-label image recognition via cross-modal fusion. In: The 29th ACM international conference on information and knowledge management, Virtual Event, Ireland, October 19-23, 2020. p. 1575–84.

[28] I. Aviles-Rivero, N. Papadakis, R. Li, S. M Alsaleh, R. T. Tan, and C.-B. Schonlieb, "When labelled data hurts: Deep semi-supervised classification with the graph 1-Laplacian," 2019, arXiv:1906.08635. [Online]. Available: http://arxiv.org/abs/ 1906.08635.

[29] AhmetIscen, GiorgosTolias, YannisAvrithis, Ondrej Chum, "Label Propagation for Deep Semi-supervised Learning", Computer Vision and Pattern Recognition, IEEE Xplore, 2019.

[30] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection", in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 936_944.

[31] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in Proc. Conf. Empirical Methods Natural Lang. Process., 2014, pp. 1532–1543.

[32] B. Chen, J. Li, X. Guo, and G. Lu, "Dualchexnet: Dual asymmetric feature learning for Thoracic disease classification in chest X-Rays," Biomed. Signal Process. Control, vol. 53, p. 101554, 2019.

[33] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition," 2014, arXiv:1402.1128.

[34] G. A. Miller, "Wordnet: A lexical database for english," Commun. The ACM, vol. 38, no. 11, pp. 39–41, 1995.

[35] J. Gao, T. Zhang, and C. Xu, "I know the relationships: Zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs," in Proc. AAAI Conf. Artif. Intell., 2019, vol. 33, pp. 8303–8311.

[36] H. Liu and P. Singh, "Conceptnet–a practical commonsense reasoning tool-kit," BT Technol. J., vol. 22, no. 4, pp. 211–226, 2004.

[37] Pennington J, Socher R, Manning CD. Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing, Doha, Qatar, a Meeting of SIGDAT, a special interest group of the ACL, October 25-29, 2014. p. 1532–43.

[38] Chen S, Chen Y, Yeh C, Wang YF. Order-free RNN With visual attention for multi-label classification. In: Proceedings of the thirty-second AAAI conference on artificial intelligence, the 30th innovative applications of artificial intelligence, and the 8th AAAI symposium on educational advances in artificial intelligence, New Orleans, Louisiana, USA, February 2-7, 2018. p. 6714–21.

[39] Chen T, Wang Z, Li G, Lin L. Recurrent attentional reinforcement learning for multi-label image recognition. In: Proceedings of the thirty-second AAAI conference on artificial intelligence, the 30th innovative applications of artificial intelligence, and the 8th AAAI symposium on educational advances in artificial intelligence, New Orleans, Louisiana, USA, February 2-7, 2018. p. 6730–7.

[40] Iscen A, Tolias G, Avrithis Y and Chum O. Label Propagation for Deep Semi-supervised Learning, Computer Vision and Pattern Recognition, 2019, pp-5070-5079. https://doi.org/10.48550/arXiv.1904.04717.