Label Co-occurrence Learning with Graph Convolutional Networks for Multi-label Chest X-ray Image Classification

Bingzhi Chen, Jinxing Li, Guangming Lu*, Hongbing Yu, and David Zhang, *Fellow, IEEE,*

***Abstract*—Existing multi-label medical image learning tasks generally contain rich relationship information among pathologies such as label co-occurrence and interdependency, which is of great importance for assisting in clinical diagnosis and can be represented as the graph-structured data. However, most state-of-the-art works only focus on regression from the input to the binary labels, failing to make full use of such valuable graph-structured information due to the complexity of graph data. In this paper, we propose a novel label co-occurrence learning framework based on Graph Convolution Networks (GCNs) to explicitly explore the dependencies between pathologies for the multi-label chest X-ray (CXR) image classification task, which we term the "CheXGCN". Specifically, the proposed CheXGCN consists of two modules, i.e., the image feature embedding (IFE) module and label co-occurrence learning (LCL) module. Thanks to the LCL model, the relationship between pathologies is generalized into a set of classifier scores by introducing the word embedding of pathologies and multi-layer graph information propagation. During end-to-end training, it can be flexibly integrated into the IFE module and then adaptively recalibrate multi-label outputs with these scores. Extensive experiments on the ChestX-ray14 and CheXpert datasets have demonstrated the effectiveness of CheXGCN as compared with the state-of-the-art baselines.**

***Index Terms*—Label co-occurrence learning; Graph Convolutional Networks; Multi-label chest X-ray image classification; Word embedding; Graph representation**

I. Introduction

CHEST X-ray (CXR) imaging [1] is the most common type of screening technique which effectively assists the clinical diagnosis and treatment for a series of thoracic diseases, such as Atelectasis, Effusion, and Pneumonia, etc. However, it is still a challenging task to assess thousands of radiology samples continuously, since their diagnosis reports rely heavily on the expert radiologists' manual annotations. More importantly, considering an individual CXR image, it might be associated with multiple abnormalities, which generally poses a considerable diagnostic challenge to the clinician. Thus, automated multi-label CXR image classification [2] has important meanings for assisting in clinical diagnosis. The purpose of automated multi-label CXR image classification is to learn a well-established classification model to help interpret any potential abnormal findings in the CXR image, so as to improve understanding and diagnostic level of CXR images. With the increased availability of large-scale CXR image datasets, various CNN-based approaches have achieved ground-breaking performance on multi-label CXR image classification. In recent years, some state-of-the-art approaches [3][2][4][5] have been proposed to address the multi-label CXR image classification task. By surveying previous works on this issue, it is noted that existing approaches can be divided into two categories: (1) Binary label approaches; (2) Label correlation approaches.

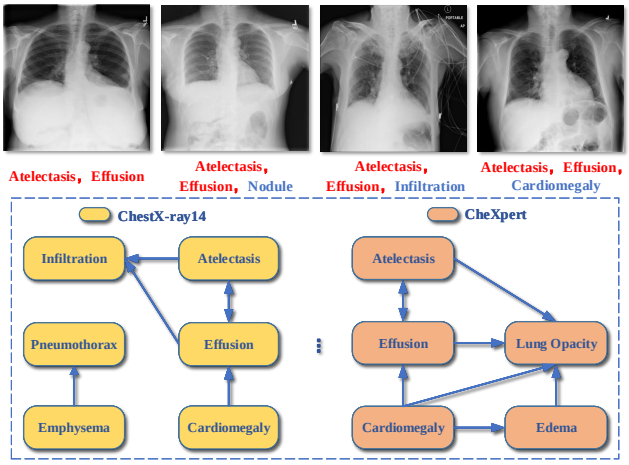


Fig. 1. Illustration of the interdependencies between labels extracted from existing chest X-ray datasets. For example, Atelectasis may also co-occur with Effusion, which can be defined as "Atelectasis → Effusion". Inspired by these interdependencies, we conduct the proposed LCL method to mine the potential labels.

In general, binary label approaches are the immediate solutions to the multi-label CXR image classification task, which transform a multi-label classification problem into multiple disjoint binary classification problems without considering any label correlations. And most previous works based on the variants of CNN models [6][7][8] can be regarded as binary relevance approaches. Although they have made tremendous efforts and achieved breakthroughs in this field, they still face some difficulties and inferiority. For each CXR image, the label annotation may be incomplete since it is unrealistic to gain a complete understanding from a complex background.

Correspondence e-mail: luguangm@hit.edu.cn

B. Chen and *G. Lu are with the Shenzhen Medical Biometrics Perception and Analysis Engineering Laboratory, Harbin Institute of Technology, Shenzhen, China. J. Li is with the School of Science and Engineering, The Chinese University of Hong Kong (Shenzhen), Shenzhen, China, and University of Science and Technology of China, Hefei, China. H. Yu is with the Nanshan District Chronic Disease Prevention and Control Hospital, Shenzhen, China. D. Zhang is with the School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, China.

Moreover, noise labels may also lead to impaired learning abilities of the binary classifier.

Fortunately, researchers have discovered that some of these abnormalities might be closely linked. Meanwhile, clinical experiences have evidenced that label co-occurrence and interdependency between abnormal patterns are important and practical significance for final diagnostics. For example, Wang et al. [2] have revealed that Infiltration tends to be linked with Atelectasis and Effusion. And Yao et al. [9] have indicated the presence of cardiomegaly is often associated with pulmonary edema. This phenomenon is not an accident in clinical diagnosis, but a result of strong correlations among pathologies, which can be called label co-occurrence [10]. In recent years, label co-occurrence has become a topic of interest to label correlation approaches. By using these interdependencies information, missing or noisy labels can be deduced from the co-occurrence relationships, which can provide radiologists with extra guidance.

Existing label correlation approaches can be generally grouped into two categories, i.e., RNN-based approaches and Attention-guided approaches. Some RNN-based approaches [9][11] are designed to utilize RNN decoder to predict the sequence of abnormalities. However, these approaches rely on the state variables to encode label correlation information and require complicated computations. Thus, they lack a robust ability to model label co-occurrences. Furthermore, recent researches [12][13][14] have favored the attention-guided approaches, which focus on the most discriminative features and channels from lesion regions in each CXR image. The main idea of these works is to model the correlations of local images (attended regions) and global images via a variety of attention mechanisms. To some extent, this strategy is considered as a popular way to implicitly capture the interdependencies between labels and attended regions. However, these methods might not be optimal since they only consider the features and channels of abnormalities extracted from the CXR images, but ignore any label co-occurrence and interdependency knowledge which do exist among different samples.

And no doubt, there still remains largely untapped information such as label co-occurrence and interdependency, which can be represented as the graph-structured data and need to be effectively leveraged, as shown in Fig. 1. Although these previous works have achieved great success on multi-label CXR image classification, they are limited by the complexity of graph-structured data in the field of graph-structured data analysis. To effectively leverage the co-occurrence information in such a knowledge graph and accurately construct the graph representation, in this paper, we propose a novel label co-occurrence learning (LCL) framework named CheXGCN, which is capable of modeling interdependency knowledge and exploring the co-occurrence of the labels based on Graph Convolutional Network (GCN) [15][16][17], greatly improving the performance of multi-label CXR image classification, as illustrated in Fig. 2.

Inspired by GCN which can encapsulate the state representation for each vertex by integrating the connectivity patterns and feature attributes from its neighbors with arbitrary depth, our method introduces the GCN-based module into the network and adaptively recalibrate multi-label prediction outputs and explore the underlying abnormities by leveraging the dependencies between different pathologies. Specifically, the proposed CheXGCN consists of two modules, i.e., the image feature embedding (IFE) module and the label co-occurrence learning (LCL) module, which play the roles of feature extractor and classifier design, respectively. The IFE module aims to learn a set of feature maps extracted from CXR images, while the LCL module is designed to generate classifiers for different pathology categories with the label embedding and multi-layer graph information propagation.

Apparently, the LCL module is the core part of the proposed CheXGCN. Here, the label embedding and structured graph representation are two key factors that influencing the effect of label co-occurrence learning. Therefore, each pathology concept in the LCL module is first represented with a semantic vector from the semantic space via word embedding. Moreover, the graph representation is represented with a predefined label correlation matrix that extracts from the co-occurrence matrix of training data, instead of learning from scratch. By feeding these semantic vectors and the existing preset correlation matrix into multiple GCN layers, we can learn a novel multi-label classifier via multi-layer graph information propagation. Finally, we optimize the whole framework via a multi-label classification loss and these generated classifiers are integrated with the image-level features extracted from the IFE module to adaptively modify the prediction beliefs for each pathology. We evaluate the effectiveness of CheXGCN on two large-scale CXR image datasets, i.e., the ChestX-ray14 [2] and CheXpert [18] dataset. And our CheXGCN reaches the state-of-the-art performance on these two datasets: the average AUC scores for 14 pathologies are 0.826 and 0.832, respectively. The main contributions of this work are summarized as follows:

(1) We present a novel GCN-based label co-occurrence learning framework (CheXGCN) that can leverage the GCN to model label co-occurrence and interdependency between different pathologies for the multi-label CXR image classification task.
(2) The proposed CheXGCN advances graph reasoning in the semantic space with multiple GCN layers, to model the interdependencies between labels and explore potential abnormalities. Moreover, the corresponding LCL module can be flexibly integrated into any CNN-based networks with end-to-end training.
(3) Experimental results on two benchmark multi-label CXR image datasets demonstrate that our CheXGCN yields superior performance over the previous state-of-the-art models.

The rest of the paper is organized as follows. Section II reviews the related works about this work. And the description of CheXGCN is given in Section III. Next, the comprehensive experiments are conducted in Section IV. Finally, Section V concludes the whole work.

II. Related Work

In this section, we make a brief summary of the related works in two aspects. On the one hand, we introduce some

previous works for automatic CXR image analysis. On the other hand, we emphasize some applications of GCNs in a variety of fields.

A. Multi-label chest X-ray image classification

With the establishment of large-scale chest X-ray datasets, the performance of multi-label chest X-ray image classification has witnessed rapid development over the past few years. Meanwhile, researchers have been made great efforts for multi-label chest X-ray image classification on some public datasets. In particular, the ChestX-ray14 dataset provided by NIH Clinical Center [1] has been the research hotspot of automatic CXR image analysis. Furthermore, other larger CXR image datasets such as CheXpert, PadChest [19] and MIMIC-CXR [20] have also been published one after another recently.

A straightforward way for this issue is to train independent binary classifiers for each abnormality with CNNs. Initially, Wang et al. [2] utilized several pretrained ImageNet [21] models to evaluate the performance of the multi-label chest X-ray image classification and found that the ResNet-50 [6] worked best. And Li et al. [22] performed thoracic disease identification and localization with additional location annotations supervision. Rajpurkar et al. [4] presented CheXNet based on a modified 121-layer DenseNet [7] model to classify abnormalities in each chest X-ray image and achieved superior performance on the detection of pneumonia. Shen et al. [23] introduced a routing-by agreement mechanism to train a CNN-based architecture for automatic thoracic disease detection. Wang et al. [24] presented a novel framework named TieNet to improve the classification performance with additional text embeddings information which was extracted from associated radiological reports. More recently, Chen at al. [5] utilized a dual asymmetric architecture based on the ResNet and DenseNet to adaptively capture more discriminative features for thoracic disease classification. However, as mentioned above, these methods ignore the relationship between labels.

In order to capture label dependencies, Yao et al. [9] presented a novel variant of DenseNet and Long-Short Term Memory Network (LSTM) [25] for multi-label chest X-ray image classification. And they preliminarily proved that exploiting label correlation can further improve the performance. Furthermore, some methods attempted to use various attention mechanisms to establish relationships between the labels and attended regions to support the global image classification. For example, Ypsilantis et al. [26] utilized an RNN-based attention model to sequentially sample the whole chest X-ray image and focus on the most informative regions. Guan et al. [27] presented a two-branch architecture that would integrate the features extracted from the global and local chest X-ray image to provide extra attention guidance for thoracic disease classification. And they later came up with a category-wise residual attention mechanism [13] to explore the correlations among pathologies, which further improve the performance of this task. Additionally, Gundel et al. [14] attempted to incorporate both high-resolution features and spatial information of pathologies to train a location-aware framework for the chest X-ray image classification. Analogously, Tang et al. [28] exploited an attention-guided curriculum learning framework that relayed the heatmaps back to the original classification task via weakly supervised localization.

[1]https://www.cc.nih.gov/drd/summers.html

B. Learning with structured graph

Another effective way to model the dependencies between labels is via using graph propagation and reasoning. For example, Marino et al. [29] applied neural networks to graph-structured data and present a GNN-based framework to mine extra attribute relationships to improve image classification. Lee et al. [30] utilized knowledge graphs to explore the correlations between multiple labels for multi-label zero-shot learning. Kipf et al. [17] presented the GCN model for graph-structured data analysis, which encoded both graph data and node features via layer-wise propagation. Based on the aforementioned GCN model, Wang et al. [31] used both semantic embeddings and the categorical relationships to predict the visual classifier for each category. Furthermore, Chen et al. [16] presented a GCN-based framework that used a pre-defined graph to model the dependencies between labels for multi-label classification.

Inspired by the aforementioned structure learning methods, we propose to learn label co-occurrence and interdependency information with GCNs to further improve the performance of multi-label CXR image classification. Compared with previous structure learning methods, our CheXGCN directly stacks multiple GCN layers to generate global classifiers for explicitly exploring the underlying abnormities. Meanwhile, the methods of label embedding and structured graph representation used in CheXGCN are more effective in dealing with the over-smoothing and scaling problems that happen with the deep depth of GCN-based architecture.

III. Label Co-occurrence Learning with CheXGCN

A. Notations and Overview

Given a dataset with c pathologies $\boldsymbol{X} = \{x_i\}_{i=1}^{c}$ where x_i represents the word embedding for m^{th} pathology. Meanwhile, each image is labeled with a c-dim label vector $\boldsymbol{L} = \{l_i\}_{i=1}^{c}$ where $l_i \in \{0, 1\}$. l_i represents whether the i^{th} pathology is presence or not, i.e., $l_i = 0$ for absence and $l_i = 1$ for presence. And the relationship graph of these pathologies is represented by $\boldsymbol{G} = (\boldsymbol{V}, \varepsilon)$ where $\boldsymbol{V}$ represents the vertex (node of pathology) and ε is the edge set that represents relationships between pathologies. Importantly, we focus on the directed graph used in each GCN layer since the relationship between pathologies tends to be one-way. For example, in Fig. 1, "Effusion $\rightarrow$ Infiltration", means when "Effusion" appears, "Infiltration" is likely to appear, but not vice versa.

In this paper, the goal of the proposed LCL method is to infer the potential label l_i with label co-occurrence and interdependency information. By using the word embedding $\boldsymbol{X}$ and the corresponding relationship graph $\boldsymbol{G}$, we utilize a GCN-based model to learn a set of classifier scores $\boldsymbol{W}$ for the image feature $\boldsymbol{F}$, in order to adaptively recalibrate the prediction beliefs for each pathology. In this way, we finally achieve more accurate label prediction outputs.

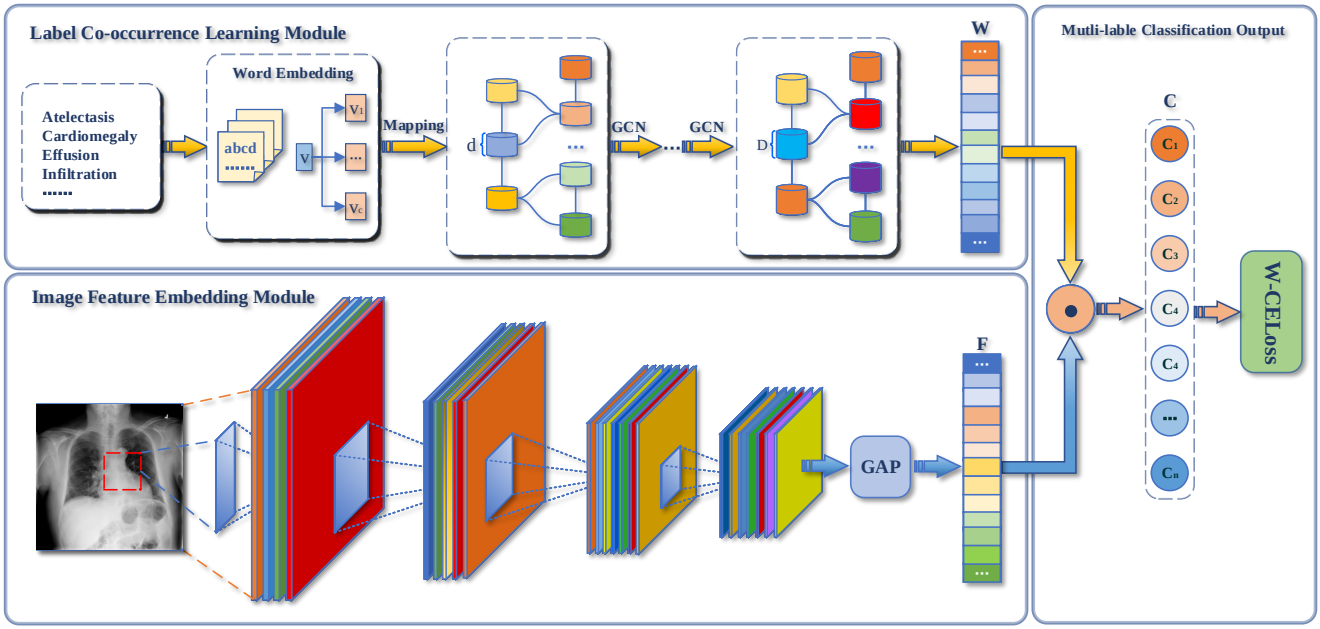


Fig. 2. Illustration of the proposed CheXGCN for multi-label CXR image classification. In the LCL module, each pathology is represented with d-dim semantic vectors $v \in \boldsymbol{V}^{n\times d}$ via word embedding, while the graph representation $\boldsymbol{G}$ is learned from the co-occurrence matrix of training data. Given the input $\boldsymbol{V}$ and $\boldsymbol{G}$ a set of classifier scores $W \in \mathbb{R}^{\boldsymbol{c}\times\boldsymbol{d}}$ are obtained by multi-step graph reasoning. In the IFE module, a basic classification subnetwork is designed to learn a set of high-level features $\boldsymbol{F} \in \mathbb{R}^{\boldsymbol{d}}$. After aggregating $\boldsymbol{W}$ and $\boldsymbol{F}$ in the end of CheXGCN, the belief states are recalibrated for the final multi-label classification outputs. In this figure, "$\odot$" represents "dot product" operation.

B. Graph Convolutional Networks

As the name might imply, the purpose of the Graph Convolutional Networks (GCNs) is to extend traditional CNN for dealing with the data represented in graph domains. And Li et al. [32] have proved that graph convolutions in GCN directly benefit from Laplacian smoothing operations. As one of the most practical graph convolutions, spectral convolution on graphs can be transformed into two parts: (1) the multiplication of a graph signal $s \in R^n$ the spectral domain; (2) the application of a spectral filter $f_{\theta'}(\theta \in R^n)$ on the spectral components. However, such a model suffers from high-cost computations. Thus, Defferrard et al. [33] presented a K-localized ChebNet for free use the spectrum and their proposed variant of convolution is defined as follows,

$$f_{\theta'}(\cdot)\mathbf{s} \approx \sum_{k=0}^{K} \theta'_k T_k(\tilde{\boldsymbol{L}})\mathbf{s} \tag{1}$$

where $f_{\theta'}$ is a function of eigenvalues of normalized graph Laplacians, $\mathbf{s}$ is the graph signal for each node, $(\cdot)$ denotes the convolution operator, $\theta' \in \mathbb{R}^K$ is the vector of Chebyshev coefficients, and T_k is the Chebyshev polynomials, as well as $\tilde{\boldsymbol{L}}$ is eigenvectors of the normalized graph Laplacian $\boldsymbol{L}$, as shown in Eq. 2 and Eq. 3.

$$\tilde{\boldsymbol{L}} = \frac{2}{\lambda_{max}}\boldsymbol{L} - \boldsymbol{I}_N \tag{2}$$

where λ_{max} denotes the largest eigenvalue of $\boldsymbol{L}$.

$$\boldsymbol{L} = \boldsymbol{I}_N - D^{-\frac{1}{2}}\boldsymbol{A}D^{-\frac{1}{2}} \tag{3}$$

where $\boldsymbol{I}_N$ is the identity matrix, $A \in \mathbb{R}(n \times n)$ is the correlation matrix, and $D = diag(d_1, d_2, \cdots, d_n)$ is the degree matrix of $\boldsymbol{A}$.

Motivated by these above works, Kipf et al. [17] proposed a multi-layer GCN model to perform semi-supervised entity classification. By approximating $\lambda_{max} = 2$ and limiting $K = 1$, the convolution operations turn to be more simplified. Moreover, they further normalize the convolution matrix $\boldsymbol{A}$ and D to tackle the vanishing gradient or exploding gradient problem in the training phase, as shown in Eq. 4.

$$\begin{aligned} f_{\theta'}(\cdot)\mathbf{s} &\approx \theta(\boldsymbol{I}_N + D^{-\frac{1}{2}}\boldsymbol{A}D^{-\frac{1}{2}}) \\ &\approx \theta(\tilde{D}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}\tilde{D}^{-\frac{1}{2}}) \end{aligned} \tag{4}$$

where $\tilde{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{I}_N$ and $\tilde{D} = \sum_j \tilde{\boldsymbol{A}}_{ij}$.

Given a structured graph with n nodes, each node is an entity represented as a d-dim feature vector $\boldsymbol{X} \in R^{n\times d}$ by word embedding. Based on the above definition of convolution, the propagation mechanism of each GCN layer can be defined as follows,

$$\boldsymbol{H}^{(l+1)} = ReLU(\tilde{D}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}\tilde{D}^{-\frac{1}{2}}\boldsymbol{H}^{(l)}\Theta^{(l)}) \tag{5}$$

where $\boldsymbol{H}^{(l)}$ is the feature descriptions of l^{th} GCN layer, and $\Theta^{(l)}$ is the corresponding trainable transformation matrix.

These simplified convolution operations can be stacked one after another with the non-linear operation (e.g. ReLU or LeakyReLU [34] operation). In our experiments, we introduce such a GCN-based model to explore the label co-occurrence and interdependency for multi-label CXR image classification.

C. Architecture of CheXGCN

To tackle the issues of automatic CXR image analysis, this paper proposes a novel framework called CheXGCN based on the design of GCN while taking into account the label

co-occurrence and interdependency. Importantly, the key to CheXGCN is to distill label co-occurrence information from the label graph with GCN, which is then applied to fine-tune the multi-label CXR image classification outputs. The architecture of CheXGCN is illustrated in Fig. 2. Specifically, two main modules (i.e., the image feature embedding (IFE) module and label co-occurrence learning (LCL) module are contained in its application architecture. Correspondingly, the IFE module is considered as the feature extractor and applied to learn the high-level features $\boldsymbol{F}$ from each CXR image, while the LCL module is built on the multiple GCN layers to learn a set of classifier scores $\boldsymbol{W}$ via multi-step graph reasoning. It is noted that the LCL module can be flexibly incorporated with different types of IFE modules in an end-to-end manner. By aggregating these corresponding scores and CNN features, CheXGCN aims to adaptively recalibrate the prediction beliefs for each pathology during the end-to-end training phase, as shown in Eq. 6.

$$\boldsymbol{P} = \boldsymbol{W} \cdot \boldsymbol{F} = \{p_i\}_{i=0}^{c} \in \mathbb{R}^c \tag{6}$$

where $\boldsymbol{P}$ is the final beliefs of pathologies. Here, the output of the LCL module is required to fit the dimensionality of the IFE module. Given the imbalance problem of the samples, a weighted Cross Entropy Loss (W-CELoss) is applied to CheXGCN for classifying the multi-label CXR image.

D. Image Feature Embedding Module

As mentioned above, the feature embedding module is initialized with the pretrained DenseNet-169 model in this work, which acts like a beginner in clinical diagnosis. The input images are first fed into the feature embedding module separately, and the discriminative features maps from the "conv_5_relu" layer are transformed by a Global average Pooling (GAP) layer to generate the initial feature representation $\boldsymbol{F} \in \mathbb{R}^{n \times d}$ of each image, as follows,

E. Label Co-occurrence Learning Module

Consistent with [16][17], the LCL module contains two-layer GCN where each layer takes the graph representation from the previous layer as the input and outputs a new graph representation. For the first GCN layer, we take the semantic-embedding vectors $\boldsymbol{X} = \{x_i\}_{i=1}^{c}$ and their corresponding relationship graph $\boldsymbol{G}$ as the input of the LCL module. According to Eq. 5, the LCL module can naturally combine graph representation and node feature in the convolution. As the key component of the proposed CheXGCN, the LCL module aims to learn a set of classifiers scores $\boldsymbol{W} \in \mathbb{R}^{c \times d}$ to recalibrate the initial beliefs for each pathology obtained from the IFE module. Above all, it is of significance to address two critical issues such as word embedding and graph representation, in order to meet the needs of the propagation mechanism based on Eq. 5.

1) Word embedding: Since the original radiology reports associated with CheX-ray14 and CheXpert datasets are not shared publicly, it is difficult to achieve a specific word embedding model for pathology embedding, which involves a large amount of corpus. Instead, we utilize the 300-dim GloVe [35] text model trained on the Wikipedia dataset to obtain the pathology embedding, to obtain the word embeddings for all pathologies in these two CXR image dataset. In our experiment, we prove that it is equally effective at conducting the proposed LCL method, as proved in Sec. IV-D and IV-E.

2) Graph representation: Obviously, the propagation rule of the node representations depends on the correlation matrix $\boldsymbol{A}$. Researchers have proposed several approaches to build up the predefined correlation matrix (explicit graph representations), instead of learning from scratch. For instance, Lee et al. [30] considered WordNet [36] as the source for building the structured knowledge graph. Gao et al. [37] used string matching to map the concepts to nodes in ConceptNet [38]. However, these strategies only use semantic embedding without any explicit interdependency information to represent relationships between labels. By contrast, we focus on the label co-occurrence matrix of training data and integrate the label co-occurrence information from adjacent nodes into a unified correlation matrix, which is applied to represent the structured graph of label relationships. The generation process is as follows:

(1) For the training data, we first count the times of appearance of pairwise pathologies $(\boldsymbol{P}_i, \boldsymbol{P}_j)$ as the elements of the label co-occurrence matrix $\boldsymbol{M}$.

$$\boldsymbol{M} = \{m_i\}_{i=1,j=0}^{c} \in \mathbb{R}^{c \times c} \tag{7}$$

(2) Based on the matrix $\mathbb{R}^{c \times c}$, we can form the initial label correlation matrix $\boldsymbol{M}'$.

$$\boldsymbol{M}' = \{m_i / N_i\}_{i=1,j=0}^{c} \in \mathbb{R}^{c \times c} \tag{8}$$

where N_i is the number of the i^{th} pathology. Thus, the graph based on such asymmetric matrix $\boldsymbol{M}'$ is a directed graph.

Obviously, the matrix $\boldsymbol{M}'$ suffers from the noise because there might exist some casual pairwise pathologies. Therefore, this paper introduces a nonlinear method for preprocessing the matrix $\boldsymbol{M}'$ that can suppress noise and protect the details of label relationships, as defined in Eq. 9.

$$\boldsymbol{A}_{i,j} = \begin{cases} 0, & \boldsymbol{M}'_{i,j} < \varphi \\ \lambda * \dfrac{\boldsymbol{M}'_{i,j}}{\sum_{j=0}^{c} \boldsymbol{M}'_j + \theta}, & \boldsymbol{M}'_{i,j} \geq \varphi \end{cases} \tag{9}$$

where $\boldsymbol{A}$ is the label correlation matrix used in each GCN layer, φ is the threshold for filtering noise, and θ is $1e-6$.

F. Structured graph propagation

Fig. 3 illustrates the structured graph propagation of the proposed LCL module. As shown, we consider each pathology as a node with states in our structured graph representation. Initially, the initial belief state $H_v^{(0)}$ is given for each pathologies node via word embedding. Then the label co-occurrence information is propagated via the learned graph representation for updating the associated belief states. The above propagation process would terminate after l steps. After propagating l times, the final belief state $H_v^{(l)}$ can be obtained to generate the multi-label classifier. Compared with the FC layer of the vanilla DenseNet model, the classifier generated by

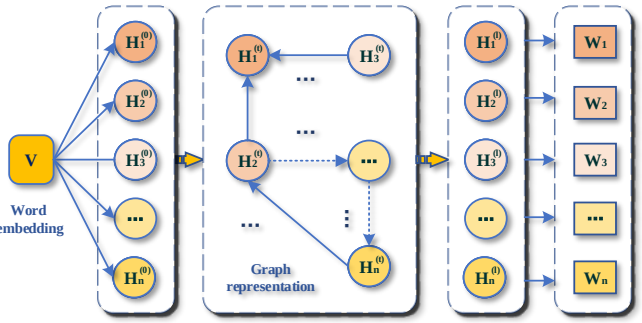


Fig. 3. Illustration of structured graph propagation with the word embedding and graph representation.

our LCL module is certainly more skillful and more efficiently in processing the potential pathologies or noisy labels. In the Section IV, we will evaluate the effect of the depth of GCNs.

G. Multi-label Classification Loss

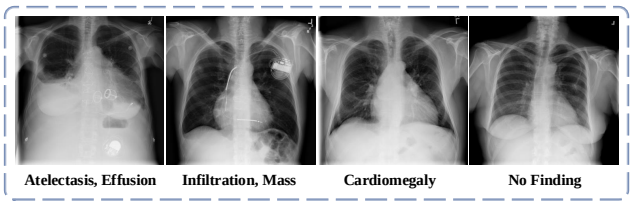
As shown in TABLE I, both ChestX-ray14 and CheXpert datasets suffer from the problem of class imbalance, which usually manifests as the imbalance between positive samples and negative samples. For example, the number of positive samples in ChestX-ray14 such as "Cardiomegaly" and "Hernia" is far less than the number of the negative sample number. To solve this problem, the proposed W-CELoss is considered as the multi-label classification loss used in our CheXGCN, which is defined as:

$$L(p_i, l_i) = -\omega_p \sum_{l_i=1} \log(\sigma(p_i)) - \omega_n \sum_{l_i=0} \log(1-\sigma(p_i)) \quad (10)$$

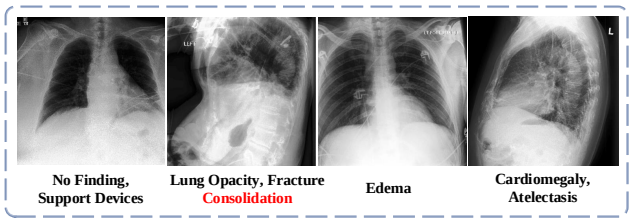
where σ is the sigmoid function, $\omega_p = \frac{|P|+|N|+|1|}{|P|+|1|}$ and $\omega_n = \frac{|P|+|N|+|1|}{|N|+|1|}$, $|P|$ and $|N|$ are the total number of positives and negatives in a batch of image labels.

TABLE I
THE NUMBER OF EACH PATHOLOGICAL CLASS IN THE CHESTX-RAY14 AND CHEXPERT DATASETS. IN THIS FIGURE, POS. AND UNC. REPRESENT THE NUMBER OF POSITIVE AND UNCERTAIN SAMPLE.

ChestX-ray14	Positive	CheXpert	Pos. (Unc.)
Atelectasis	11,559	No Finding	16,627 (0)
Cardiomegaly	2,776	En_Cardiom.	9,020 (10,148)
Effusion	13,317	Cardiomegaly	23,002 (6,597)
Infiltration	19,894	Lung Lesion	6,856 (1,071)
Mass	5,782	Lung Opacity	92,669 (4,341)
Nodule	6,331	Edema	48,905 (11,571)
Pneumonia	1,432	Consolidation	12,730 (23,976)
Pneumothorax	5,302	Pneumonia	4,576 (15,658)
Consolidation	4,667	Atelectasis	29,333 (29,377)
Edema	2,303	Pneumothorax	17,313 (2,663)
Emphysema	2,516	Effusion	75,696 (9,419)
Fibrosis	1,686	Pleural Other	2,441 (1,771)
P_T	3,385	Fracture	7,270 (484)
Hernia	227	Devices	105,831 (898)



(a) ChestX-ray14



(b) CheXpert

Fig. 4. Illustration of example images with their corresponding pathologies, which are selected from the ChestX-ray14 (a) and CheXpert (b) datasets, respectively. In this figure, the uncertain pathology is marked in red.

IV. EXPERIMENTS

In this section, the proposed CheXGCN is evaluated on the ChestX-ray14 and CheXpert datasets. We first give descriptions of the datasets. Next, we go into details of implementation details. Then we evaluate the LCL method and give detailed experimental analyses. In addition, the benefits brought by the LCL module are verified through an extensive ablation study. Comparison with the state-of-the-art methods is presented at last.

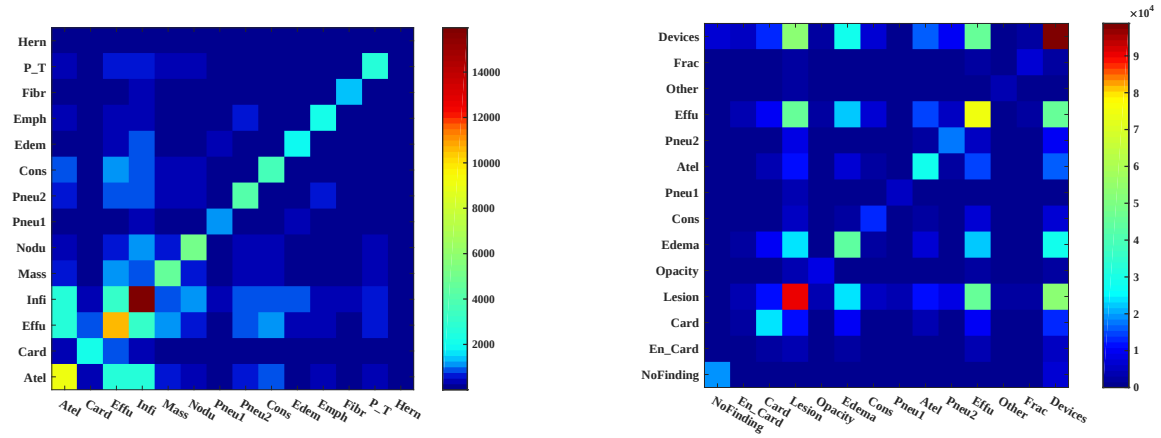
A. Datasets

In recent years, most previous works focus on the ChestX-ray14 dataset and have made great efforts in the promotion of automatic CXR image analysis. Moreover, an extra large-scale CXR image dataset called CheXpert is also considered as the evaluation benchmark. All the definitions of the observations used in these two datasets are conforming to the Fleischner Society's recommended glossary [1] whenever applicable.

1) ChestX-ray14: The ChestX-ray14[2] dataset is a widely used benchmark for multi-label CXR image classification. It consists of 112,120 frontal-view X-ray images with 14 common disease pathologies. In ChestX-ray14, each image is labeled with one or more pathologies or "No Finding" (if there is no any abnormity in an image), as illustrated in Fig. 4(a).

2) CheXpert: The CheXpert[3] dataset is another large-scale public benchmark for chest radiograph interpretation. It consists of 224,316 multi-view CXR images (i.e., the frontal and lateral images) which labeled for the presence of 14 observations as positive, negative, or uncertain, as illustrated in Fig. 4(b). Importantly, the uncertain labels in CheXpert could

[2]ChestX-ray14: https://nihcc.app.box.com/v/ChestXray-NIHCC
[3]CheXpert: https://stanfordmlgroup.github.io/competitions/chexpertl



(a) ChestX-ray14 (b) CheXpert

Fig. 5. Illustration of label co-occurrence matrixes extracted from the ChestX-ray14 (a) and CheXpert (b) datasets, respectively.

be regarded as positives or negatives. Compared with ChestX-ray14, CheXpert has a larger scale and higher label accuracy. Note that, the test set of CheXpert has not been revealed yet.

As mentioned, the proposed LCL method refers to the problem of assigning potential labels to each object according to the co-occurring label pairs. Here, both the ChestX-ray14 and CheXpert are the multi-label CXR datasets and exist a large number of the co-occurring label pairs, as shown in Fig. 5. For example, Atelectasis often occurs with Infiltration, while Lung Lesion is often associated with Pleural Effusion. This phenomenon preliminary proves the feasibility of the proposed LCL methods. Furthermore, it also indicates that both two datasets suffer from the class imbalance problem. For example, in ChestX-ray14, the occurrence of "Infiltration" far exceeded other pathologies.

B. Implementation Details

Normally most previous works randomly shuffle the ChestX-ray14 dataset into three subsets: 70% for training, 10% for validation and 20% for testing. However, as mentioned in [14], there are remarkable differences in experimental results between different test sets because of the class imbalance problem. To make our results directly comparable to most published baselines, we strictly follow the official split standards of ChestX-ray14 provided by Wang et al. [2] to conduct our experiments. Due to the test set of the CheXpert dataset is not yet available for open publication, we also randomly shuffle the existing images into three subsets: 70% for training, 10% for validation and 20% for testing. Note that, no uncertain label is present in our validation and test sets. Following the uncertainty approaches in [18], two distinct ways are applied to explicitly incorporate the uncertainty labels in CheXpert, which we term CheXpert_1s and CheXpert_0s. In detail, CheXpert_1s considers uncertainty labels as positive labels. Instead, uncertainty labels for CheXpert_0s are processed to the negative labels.

Futhermore, the IFE module in CheXGCN is initialized with the pretrained DenseNet-169 model. And the LCL module is composed of two GCN layers with output channel numbers as $300 \rightarrow 1024 \rightarrow d$, where $d = 1664$ represents the dimensionality of the pathology classifier. In addition, we consider LeakyReLU with the negative slope of 0.2 as the non-linear activation function used in the LCL module. Furthermore, the proposed approach is conducted by using the deep learning toolbox PyTorch [39] and runs on two Nvidia Titan XP GPU with 12 GB memory. In our experiments, the mini-batch size is set to 8 and the initial learning rate is set to 0.001, which is decreased 10 times every 5 epochs. Moreover, we first resize the original images to 556×556 pixels, then randomly crop it to 512×512 pixels. During training, we use stochastic gradient descent (SGD) [40] with 0.9 momentum. Consistent with the aforementioned baselines, we consider the AUC score [41] (the area under the ROC curve [41]) as the evaluation metric to evaluate the proposed CheXGCN.

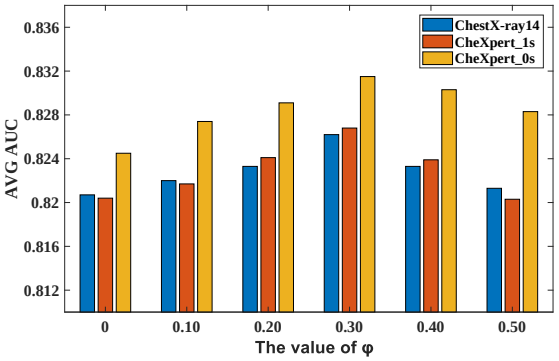
C. Parameter Analysis

In this section, we mainly perform evaluations of parameter analysis from four three aspects, including the sensitivity of parameters φ and λ in Eq. 9, and the effect of the depth of GCNs for the multi-label CXR image classification task.

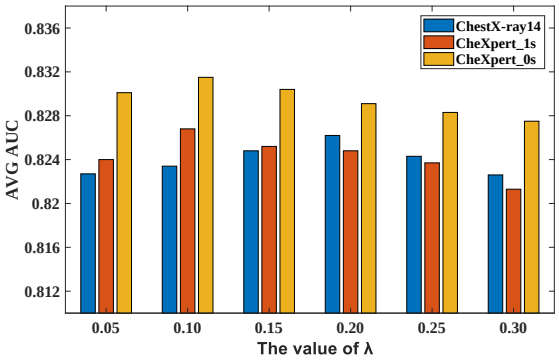
1) Evaluation on threshold value φ: The threshold value φ in Eq. 9 is one of the key parameters for filtering the noisy edges of graph representation. By fixing other parameters, we conduct the proposed CheXGCN with a range of different threshold values φ, i.e., $\varphi \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$. As shown in Fig. 6(a), the proposed CheXGCN achieves the best performance on both ChestX-ray14 and CheXpert when the value of φ is set to be 0.3.

2) Evaluation on parameter λ: The parameter λ in Eq. 9 is designed to control the correlation state between the node and its neighborhood, and has important impact on graph representation. Similarly, by fixing other parameters, we also evaluate a range of different values of parameter λ, i.e., $\lambda \in \{0.05, 0.10, 0.15, 0.25, 0.30\}$. As shown in Fig. 6(b), the optimal value of λ is 0.2 for ChestX-ray14, while it is 0.3 for CheXpert_1s and 0.1 for CheXpert_0s.

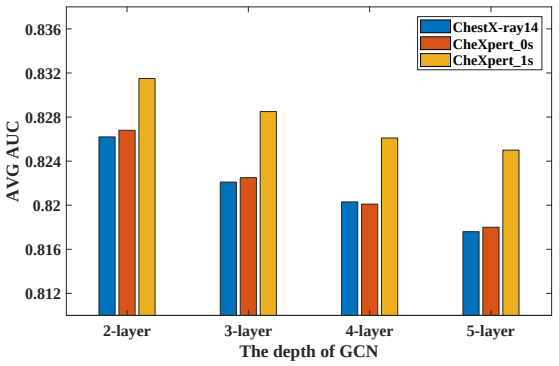
3) Evaluation on the depth of GCNs: The depth of GCN is an important issue for the proposed CheXGCN. In our



(a) Sensitivity of parameters φ



(b) Sensitivity of parameters λ



(c) Effect of the depth of GCNs

Fig. 6. Comparison of average AUC scores for all 14 pathologies with different values of φ, λ and the depth of GCNs, respectively.

experiments, we evaluate an m-layer GCN for node information propagation, where $m \in \{2, 3, 4, 5\}$. Specifically, for the 3-layer model, we consider the LCL module with output channel numbers as $300 \rightarrow 1024 \rightarrow 1024 \rightarrow 1664$. And the 4-layer model has output channel numbers as $300 \rightarrow 1024 \rightarrow 1024 \rightarrow 1024 \rightarrow 1664$. Moreover, the output channel numbers of 5-layer model are set to $300 \rightarrow 1024 \rightarrow 1024 \rightarrow 1024 \rightarrow 1024 \rightarrow 1664$. As shown in Fig. 6(c), the proposed CheXGCN achieves the best performance of both ChestX-ray14 and CheXpert when we consider a 2-layer GCN for label co-occurrence learning, and the performance keeps decreasing as we make the LCL module deeper from 2-layer to 5-layer. We suspect that this is because node representation tends to converge when using more GCN layers.

D. CheXGCN vs. CNN model

Based on the above parameter settings, we compute the AUC score of each class and the average AUC scores across all pathologies. Quantitative results are reported in TABLE II and TABLE III. To show the effectiveness of the proposed LCL method, we first compare our CheXGCN with the corresponding CNN model (i.e., the vanilla DenseNet model with the FC layer) on the ChestX-ray14 and CheXpert datasets, respectively.

1) Results on ChestX-ray14: As shown in TABLE II, the proposed CheXGCN comprehensively improves the multi-label classification performance of ChestX-ray14 and its overall performance outperforms the vanilla DenseNet-169 model by 1.2% (0.826 vs. 0.814). Particularly the AUC scores of CheXGCN for 4 pathologies surpass the DenseNet-169 over 1.5%, i.e., Mass (0.840 vs. 0.824), Consolidation (0.751 vs. 0.735), as well as Fibrosis (0.819 vs. 0.834) and Hernia (0.903 vs. 0.929). Furthermore, it is obvious to see that the AUC scores of the Infiltration group are far lower than others. It is because Infiltration is often difficult to identify due to its various textures. By contrast, some pathologies such as Cardiomegaly, Emphysema, and Hernia are more likely to be diagnosed accurately since they generally have specific pathological characteristics.

2) Results on CheXpert: As shown in TABLE III, both CheXGCN_1s and CheXGCN_0s perform better than the corresponding DenseNet-169 on the CheXpert dataset and their average AUC scores are improved 0.9% and 0.7% (0.827 vs. 0.818 & 0.832 vs. 0.823), respectively. Results show the proposed LCL method can obviously meliorate the diagnosis effects of some pathologies and improves their AUC scores, such as Lung Opacity, Pneumothorax and Pleural Other, etc. Overall, we can found that CheXGCN_0s outperforms CheXGCN_1s (0.832 vs. 0.827), especially for the detection of Enlarged Cardiomediastinum (0.697 vs. 0.682), Consolidation (0.784 vs. 0.745) and Pleural Other (0.835 vs. 0.823). This was because the uncertainty labels of these pathologies are likely to be negative, and vice versa.

E. CheXGCN vs. State-of-the-Art baselines

In this part, some state-of-the-art baselines for multi-label CXR image classification on these two datasets are applied to the comparative experiments. On the one hand, our CheXGCN is compared to a variety of methods based on the ChestX-ray14 dataset, including U-DCNN (Wang et al. [2]), LSTM-Net (Yao et al. [9]), CheXNet (Rajpurkar et al. [4]), DNet (Gundel et al. [14]), AGCL (Tang et al. [28]), DR-DNN (Shen et al. [23]), CRAL (Guan et al. [13]) as well as DualCheXN (Chen et al. [5]). On the other hand, we further evaluate our model as compared to the original uncertainty approaches (U_Ones and U_Zeros) [18] based on the CheXpert dataset.

1) Results on ChestX-ray14: As shown in TABLE II, the proposed CheXGCN contributes a new state-of-the-art: it yields an average AUC score of 0.826 for all 14 pathologies and achieves the top performance for more than half of pathologies. Fig. 7(a) illustrates the ROC curves of

TABLE II
COMPARISON WITH PREVIOUS BASELINES ON THE CHESTX-RAY14 DATASET. THE ACU SCORES OF EACH PATHOLOGY IN CHESTX-RAY14 ARE REPORTED. FOR EACH COLUMN, THE BEST RESULTS IS HIGHLIGHTED IN **BOLD**.

Method	Atel	Card	Effu	Infi	Mass	Nodu	Pneu1	Pneu2	Cons	Edem	Emph	Fibr	P_T	Hern	**Mean**
U-DCNN [2]	0.700	0.810	0.759	0.661	0.693	0.669	0.658	0.799	0.703	0.805	0.833	0.786	0.684	0.872	0.745
LSTM-Net[9]	0.733	0.856	0.806	0.673	0.718	0.777	0.684	0.805	0.711	0.806	0.842	0.743	0.724	0.775	0.761
DR-DNN [23]	0.766	0.801	0.797	**0.751**	0.760	0.741	**0.778**	0.800	**0.787**	0.820	0.773	0.765	0.759	0.748	0.775
AGCL [28]	0.756	0.887	0.819	0.689	0.814	0.755	0.729	0.850	0.728	0.848	0.906	0.818	0.765	0.875	0.803
CheXNet [4]	0.769	0.885	0.825	0.694	0.824	0.759	0.715	0.852	0.745	0.842	0.906	0.821	0.766	0.901	0.807
DNet [14]	0.767	0.883	0.828	0.709	0.821	0.758	0.731	0.846	0.745	0.835	0.895	0.818	0.761	0.896	0.807
CRAL [13]	0.781	0.880	0.829	0.702	0.834	0.773	0.729	0.857	0.754	0.850	0.908	0.830	0.778	0.917	0.816
DualCheXN[5]	0.784	0.888	0.831	0.705	0.838	0.796	0.727	0.876	0.746	**0.852**	0.942	**0.837**	**0.796**	0.912	0.823
DenseNet-169	0.778	0.879	0.826	0.697	0.824	0.792	0.726	0.867	0.735	0.841	0.932	0.819	0.782	0.903	0.814
CheXGCN	**0.786**	**0.893**	**0.832**	0.699	**0.840**	**0.800**	0.739	**0.876**	0.751	0.850	**0.944**	0.834	0.795	**0.929**	**0.826**

∗The 14 pathologies in ChestX-ray14 are Atelectasis (Atel), Cardiomegaly (Card), Effusion (Effu), Infiltration (Infi), Mass, Nodule (Nodu), Pneumonia (Pneu1), Pneumothorax (Pneu2), Consolidation (Cons), Edema (Edem), Emphysema (Emph), Fibrosis (Fibr), Pleural Thickening (P_T) and Hernia (Hern).

TABLE III
COMPARISON WITH PREVIOUS BASELINES ON THE CHEXPERT DATASET. THE ACU SCORES OF EACH PATHOLOGY IN CHEXPERT ARE REPORTED.. FOR EACH COLUMN, THE BEST RESULTS IS HIGHLIGHTED IN **BOLD**.

(a) CheXGCN_1s

Method	NoFi	EnCa	Card	Lesi	Opac	Edem	Cons	Pneu1	Atel	Pneu2	Effu	Other	Frac	Devi	**Mean**
U_Ones [18]	0.875	0.676	0.873	0.764	0.795	0.880	0.735	0.794	0.722	0.898	0.901	0.805	0.791	0.896	0.815
DenseNet_169	0.877	0.678	**0.877**	0.767	0.803	0.882	0.736	0.792	0.724	0.903	0.902	0.805	0.811	0.898	0.818
CheXGCN_1s	**0.879**	**0.682**	0.876	**0.768**	**0.821**	**0.884**	**0.745**	**0.805**	**0.731**	**0.913**	**0.906**	**0.823**	**0.842**	**0.902**	**0.827**

(b) CheXGCN_0s

Method	NoFi	EnCa	Card	Lesi	Opac	Edem	Cons	Pneu1	Atel	Pneu2	Effu	Other	Frac	Devi	**Mean**
U_Zeros [18]	0.877	0.691	0.876	0.766	0.807	0.881	0.775	0.804	0.730	0.902	0.902	0.815	0.807	0.893	0.823
DenseNet_169	0.878	0.692	0.876	0.764	0.809	0.882	0.778	0.800	0.732	0.906	0.902	0.817	0.824	0.894	0.825
CheXGCN_0s	**0.879**	**0.697**	**0.877**	**0.768**	**0.822**	**0.886**	**0.784**	**0.810**	**0.736**	**0.917**	**0.907**	**0.835**	**0.833**	**0.899**	**0.832**

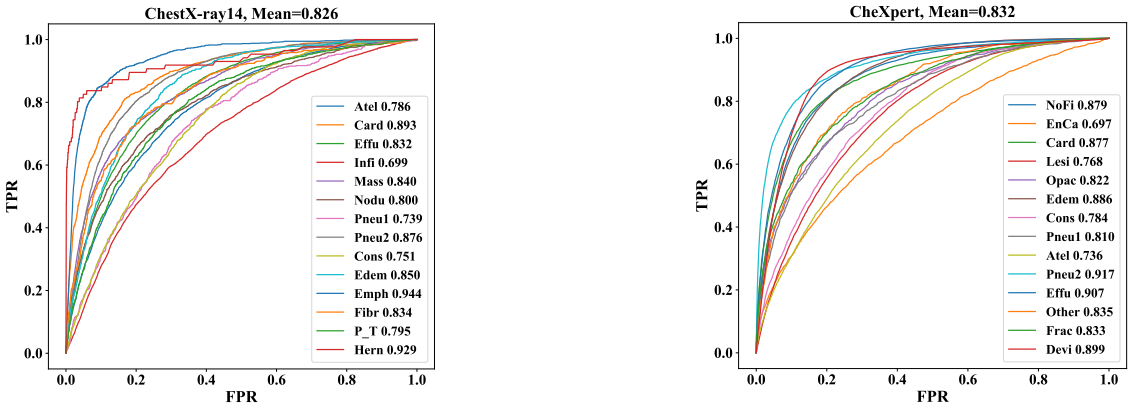
∗The 14 pathologies in CheXpert are No Finding (NoFi), Enlarged Cardiomediastinum (EnCa), Cardiomegaly (Card), Lung Lesion (Lesi), Lung Opacity (Opac), Edema (Edem), Consolidation (Cons), Pneumonia (Pneu1), Atelectasis (Atel), Pneumothorax (Pneu2), Pleural Effusion (Effu), Pleural Other (Other), Fracture (Frac) and Support Devices (Devi), respectively.

CheXGCN over the 14 pathologies on ChestX-ray14. Specifically, CheXGCN outperforms these previous baselines, especially for U-DCNN (0.745) and LSTM-DNet (0.761) with improvements of 8.1% and 6.5%. Moreover, the overall performance of CheXGCN has an improvement of 1% over CRAL (0.816). Compared with CRAL, the AUC scores of some pathologies with CheXGCN are obviously improved, e.g. Cardiomegaly (0.893 vs. 0.880), Nodule (0.800 vs. 0.773), and Emphysema (0.944 vs. 0.908) and Pleural Thickening (0.795 vs. 0.778). Importantly, the proposed CheXGCN outperforms DualCheXNet (0.826 vs. 0.823) and improves the classification performance of detecting Pneumonia (0.739 vs. 0.729) and Hernia (0.929 vs. 0.912) by more than 1%. This improvement is a satisfactory outcome because the texture of pneumonia is variegated and hard-to-identify.

2) Results on CheXpert: As shown in TABLE III, both CheXGCN_1s and CheXGCN_0s achieve higher average AUC scores for all 14 pathologies in CheXpert as compared to U_Ones and U_Zeros. And our CheXGCN_0s contributes a new state-of-the-art and surpasses the previous state-of-the-art (i.e., U_Zeros) by 0.9% (0.832 vs. 0.823). Fig. 7(b) illustrates the ROC curves of CheXGCN over the 14 pathologies on CheXpert. In particular, the AUC scores of some pathologies such as Lung Opacity (0.822 vs .0.807), Pneumothorax (0.917 vs. 0.902), as well as Pleural Other (0.835 vs. 0.815) and Fracture (0.833 vs 0.807) are improved by more than 1.5% as compared to the current highest scores.

F. Robustness of CheXGCN

On the bases of the experimental results stated above, the improvements of the proposed CheXGCN meet the ChestX-ray14 dataset for 'statistical significance', and not disappear when the test set is repeated with other different samples selected from the CheXpert dataset. The classification performance of the proposed CheXGCN compares favorably against the corresponding CNN model and the previous baselines on both ChestX-ray14 and CheXpert datasets, confirming its robustness in the field of multi-label CXR image analysis.



(a) ChestX-ray14 (b) CheXpert

Fig. 7. ROC curves of CheXGCN on the ChestX-ray14 (a) and CheXpert (b) datasets, respectively. The corresponding AUC scores are given in TABLE II and TABLE III.

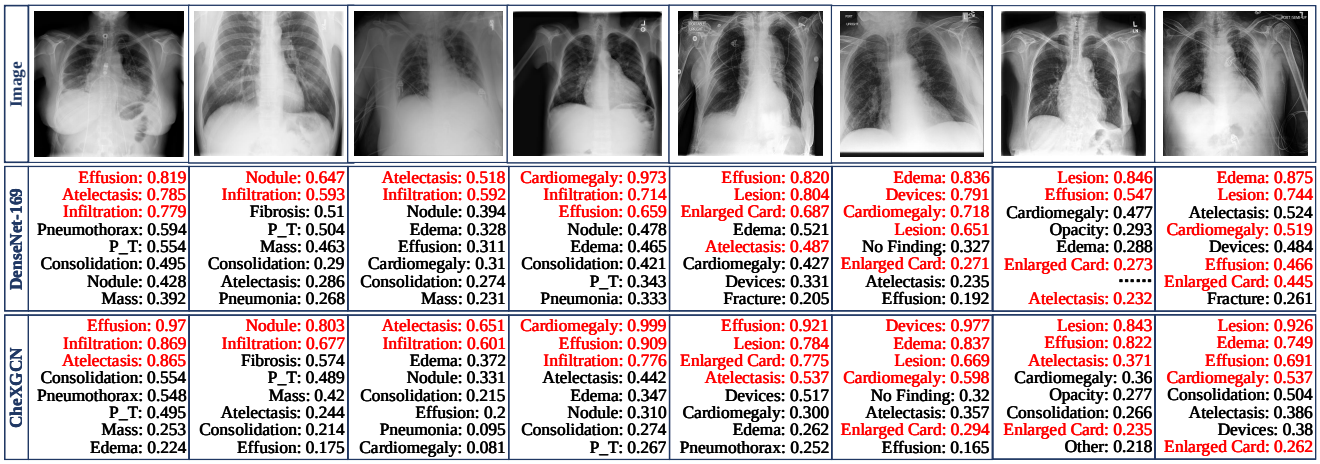


Fig. 8. Test results of the multi-label CXR image classification with CheXGCN on the ChestX-ray14 and CheXpert datasets. The top-8 predicted categories and their corresponding probability scores are presented. In this figure, the ground truth pathologies are highlighted in red

G. Visualization Results

Fig. 8 illustrates the intuitive presentations of multi-label CXR image classification. The top-8 prediction scores are presented for each test image with the highest values at the top. Compared with the vanilla DenseNet-169 model, the proposed CheXGCN can surely enhance the effects of multi-label CXR image classification. With the full consideration of the label-wise relationships, our CheXGCN can not only effectively improve the confidence scores of ground truth pathologies but also reduce the probabilities of other irrelevant labels. For example, in column 7, in the condition of Effusion appearing, CheXGCN increases the corresponding scores of Atelectasis ($0.232 \rightarrow 0.371$) since it is often associated with Effusion.

V. Conclusion

Label co-occurrence learning is of great importance in the field of multi-label medical image analysis. However, it is still a challenge to effectively capture the label dependencies in the field of multi-label medical image analysis. In this paper, a novel GCN-based architecture named CheXGCN is proposed to leverage label co-occurrence and interdependency information for improving the multi-label CXR image classification task. By incorporating the word embedding of the pathologies and their graph representation, the proposed CheXGCN aims to learn a set of classifier scores to fine-tune the prediction belief states for each pathology via multi-layer graph information propagation. In our experiments, CheXGCN achieves a satisfactory performance of multi-label CXR image classification on both ChestX-ray14 and CheXpert datasets. More importantly, we also demonstrate the competitive generalization performance of CheXGCN by comparing with the state-of-the-art approaches.

Acknowledgment

The work was supported in part by the National Natural Science Foundation of China Fund under Grant 61906162, in part by the Shenzhen Fundamental Research Fund under Grants JCYJ20180306172023949 and JCYJ20170811155442454, in part by the China Postdoctoral Science Foundation under Grants 2019TQ0316 and 2019M662198, in part by the Medical Biometrics Perception and Analysis Engineering Labora-

tory, Shenzhen, China, and in part by the Shenzhen Research Institute of Big Data, in part by the Shenzhen Institute of Artificial Intelligence and Robotics for Society.

REFERENCES

[1] D. M. Hansell, A. A. Bankier, H. MacMahon, T. C. McLoud, N. L. Muller, and J. Remy, "Fleischner society: glossary of terms for thoracic imaging," *Radiology*, vol. 246, no. 3, pp. 697–722, 2008.

[2] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2097–2106.

[3] H. Salehinejad, E. Colak, T. Dowdell, J. Barfett, and S. Valaee, "Synthesizing chest x-ray pathology for training deep convolutional neural networks," *IEEE transactions on medical imaging*, vol. 38, no. 5, pp. 1197–1206, 2018.

[4] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya *et al.*, "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *CoRR*, vol. abs/1711.05225, 2017.

[5] B. Chen, J. Li, X. Guo, and G. Lu, "Dualchexnet: dual asymmetric feature learning for thoracic disease classification in chest x-rays," *Biomedical Signal Processing and Control*, vol. 53, p. 101554, 2019.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[7] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[8] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[9] L. Yao, E. Poblenz, D. Dagunts, B. Covington, D. Bernard, and K. Lyman, "Learning to diagnose from scratch by exploiting dependencies among labels," *CoRR*, vol. abs/1710.10501, 2017.

[10] C. Galleguillos, A. Rabinovich, and S. Belongie, "Object categorization using co-occurrence, location and appearance," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.

[11] H.-C. Shin, K. Roberts, L. Lu, D. Demner-Fushman, J. Yao, and R. M. Summers, "Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2497–2506.

[12] J. Cai, L. Lu, A. P. Harrison, X. Shi, P. Chen, and L. Yang, "Iterative attention mining for weakly supervised thoracic disease pattern localization in chest x-rays," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 589–598.

[13] Q. Guan and Y. Huang, "Multi-label chest x-ray image classification via category-wise residual attention learning," *Pattern Recognition Letters*, 2018, doi:10.1016/j.patrec.2018.10.027.

[14] S. Guendel, S. Grbic, B. Georgescu, S. Liu, A. Maier, and D. Comaniciu, "Learning to recognize abnormalities in chest x-rays with location-aware dense networks," in *Iberoamerican Congress on Pattern Recognition*. Springer, 2018, pp. 757–765.

[15] J. Zhou, G. Cui, Z. Zhang, C. Yang, Z. Liu, and M. Sun, "Graph neural networks: A review of methods and applications," *CoRR*, vol. abs/1901.00596, 2018.

[16] Z.-M. Chen, X.-S. Wei, P. Wang, and Y. Guo, "Multi-label image recognition with graph convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5177–5186.

[17] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *International Conference on Learning Representations*, 2016.

[18] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya *et al.*, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Thirty-Third AAAI Conference on Artificial Intelligence*, 2019.

[19] A. Bustos, A. Pertusa, J.-M. Salinas, and M. de la Iglesia-Vayá, "Padchest: A large chest x-ray image dataset with multi-label annotated reports," *CoRR*, vol. abs/1901.07441, 2019.

[20] P. Rajpurkar, J. Irvin, A. Bagul, D. Ding, T. Duan, H. Mehta, B. Yang, K. Zhu, D. Laird, R. L. Ball *et al.*, "Mura: Large dataset for abnormality detection in musculoskeletal radiographs," *CoRR*, vol. abs/1712.06957, 2017.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[22] Z. Li, C. Wang, M. Han, Y. Xue, W. Wei, L.-J. Li, and L. Fei-Fei, "Thoracic disease identification and localization with limited supervision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8290–8299.

[23] Y. Shen and M. Gao, "Dynamic routing on deep neural network for thoracic disease classification and sensitive area localization," in *International Workshop on Machine Learning in Medical Imaging*. Springer, 2018, pp. 389–397.

[24] X. Wang, Y. Peng, L. Lu, Z. Lu, and R. M. Summers, "Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9049–9058.

[25] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition," *CoRR*, vol. abs/1402.1128, 2014.

[26] P.-P. Ypsilantis and G. Montana, "Learning what to look in chest x-rays with a recurrent visual attention model," *CoRR*, vol. abs/1701.06452, 2017.

[27] Q. Guan, Y. Huang, Z. Zhong, Z. Zheng, L. Zheng, and Y. Yang, "Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification," *CoRR*, vol. abs/1801.09927, 2018.

[28] Y. Tang, X. Wang, A. P. Harrison, L. Lu, J. Xiao, and R. M. Summers, "Attention-guided curriculum learning for weakly supervised classification and localization of thoracic diseases on chest radiographs," in *International Workshop on Machine Learning in Medical Imaging*. Springer, 2018, pp. 249–258.

[29] K. Marino, R. Salakhutdinov, and A. Gupta, "The more you know: Using knowledge graphs for image classification," *Conference on Computer Vision and Pattern Recognition*, pp. 20–28, 2016.

[30] C.-W. Lee, W. Fang, C.-K. Yeh, and Y.-C. Frank Wang, "Multi-label zero-shot learning with structured knowledge graphs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1576–1585.

[31] X. Wang, Y. Ye, and A. Gupta, "Zero-shot recognition via semantic embeddings and knowledge graphs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6857–6866.

[32] Q. Li, Z. Han, and X.-M. Wu, "Deeper insights into graph convolutional networks for semi-supervised learning," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[33] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," in *Advances in neural information processing systems*, 2016, pp. 3844–3852.

[34] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. icml*, vol. 30, no. 1, 2013, p. 3.

[35] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[36] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[37] J. Gao, T. Zhang, and C. Xu, "I know the relationships: Zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs," 2019.

[38] H. Liu and P. Singh, "Conceptnet—a practical commonsense reasoning tool-kit," *BT technology journal*, vol. 22, no. 4, pp. 211–226, 2004.

[39] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[40] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*. Springer, 2010, pp. 177–186.

[41] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, and M. Müller, "proc: an open-source package for r and s+ to analyze and compare roc curves," *BMC bioinformatics*, vol. 12, no. 1, p. 77, 2011.