UNIVERSITY OF TECHNOLOGY SYDNEY
Faculty of Engineering and Information Technology

CHEST X-RAY IMAGE CLASSIFICATION WITH DEEP LEARNING

by

Qingji Guan

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

Sydney, Australia

2021

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Qingji Guan, declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise reference or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Production Note:
Signature removed prior to publication.

Signature:

Date: 1 Sep 2021

ABSTRACT

CHEST X-RAY IMAGE CLASSIFICATION WITH DEEP LEARNING

by

Qingji Guan

Computer-aided diagnosis (CAD) systems have been successfully helped to clinical diagnosis. This dissertation considers one essential task in CAD, the chest X-ray (CXR) image classification problem, with the deep learning technologies from the following three aspects.

First, considering most diseases existing in CXRs usually happen in small localized areas, we propose to localize the local discriminative regions and integrate the global and local cues into an attention guided convolution neural network (AG-CNN) to identify thorax diseases. AG-CNN consists of three branches (global, local, and fusion branches). The global branch learns the global features for classification. The local branch localizes the discriminative regions, which avoids noise and improves misalignment in the global branch. AG-CNN fuses the global and local features for diagnosis in a fusion branch.

Second, due to the common and complex relationships of multiple diseases in CXRs, it is worth exploiting their correlations to help the diagnosis. This thesis will present a category-wise residual attention learning method to concentrate on learning the correlations of multiple diseases. It is expected to suppress the obstacles of irrelevant categories and strengthen the relevant features at the same time.

Last, a robust and stable CXR image analysis system should be able to: 1) automatically focus on the disease-critical regions, which usually are of small sizes; 2) adaptively capture the intrinsic relationships among different disease features and utilize them to boost the multi-label disease recognition rates jointly. We introduce a discriminative feature learning framework, ConsultNet, to achieve those two purposes simultaneously. ConsultNet consists of a variational selective information bottleneck branch and a spatial-

and-channel encoding branch. These two branches learn discriminative features collaboratively.

In addition, each of the proposed methods is comprehensively verified and analyzed by conducting various experiments.

Dissertation directed by Professor Yi Yang
Australian Artificial Intelligence Institute (AAII), School of Computer Science

Acknowledgements

I would like to dedicate this dissertation to those who have offered assistance and support while pursuing the Ph.D. degree.

First and foremost, I would like to thank my supervisor, Prof. Yi Yang, for his guidance, encouragement, and most of all, his support during my preparation for the thesis. Prof. Yang taught me to insist on being myself and always do the right things in research work. He has also given me many valuable suggestions in lifetime. He has my deepest gratitude.

Besides, I would like to thank my co-supervisor, Dr. Liang Zheng, who has given me valuable suggestions and guidance in my research. His profound thinking and painstaking inspire and encourage me to do my best to do perfect work.

Thanks to all my colleagues in the ReLER Lab at the University of Technology Sydney. Thanks to Qianyu Feng, Peike Li, Pingbo Pan, Zhedong Zheng, Fan Ma, Xiaohan Wang, Yu Wu, Yutian Lin, Hehe Fan, Yanbin Liu, Jiaxu Miao, Hu Zhang, Guang Li, Xiaolin Zhang, Ruoyu Liu, Yang He, Zongxin Yang. Special thanks to Zhun Zhong, Yawei Luo and Ping Liu, who have many discussions with my research together. I am grateful and delighted to meet these lovely, passionate people.

I would like to thank my parents Fengxian Sheng and Chengwen Guan, my brother Qingli Guan for their love over the years.

Finally and most importantly, I would like to thank my husband, Liming Yang, for his selfless love. Without his tremendous support, I could not complete my study. I thank my son, Yichen Yang, who always cheers me up.

Qingji Guan
March 2021 at Beijing.

List of Publications

Journal Papers

J-1. **Q. Guan**, Y. Huang, Y. Luo, P. Liu, M. Xu and Y. Yang, “Discrimative feature learning for thorax disease classification in chest X-ray images,” *IEEE Transactions on Image Processing*, vol. 30, 2021.

J-2. **Q. Guan**, Y. Huang, Z. Zhong, Z. Zheng, L. Zheng and Y. Yang. “Thorax disease classification with attention guided convolutional neural network ,” *Pattern Recognition Letters*, vol. 131, 38-45, 2020.

J-3. **Q. Guan** and Y. Huang. “Multi-label chest X-ray image classification via category-wise residual attention learning ,” *Pattern Recognition Letters*, vol. 130, 259-266, 2020.

Contents

List of Figures

Chapter 1

Introduction

Chest X-ray (CXR) image classification is one of the most important tasks in computer-aided thorax disease diagnosis. The clinical diagnosis of thorax diseases mainly relies on professional knowledge and careful manual observation. Due to the complex pathologies and subtle texture changes of different lung lesions in images, radiologists may make mistakes even though they have experienced long-term clinical training and professional guidance. Therefore, it is very important to develop the CXR image classification methods to support clinical practitioners.

1.1 Background

CXR images own their particular characteristics compared with the general natural images (*e.g.*, ImageNet (Deng et al., 2009)) in the following aspects.

1. The lesion areas are relatively small compared with the entire image. A large amount of healthy areas or disease-irrelevant regions would introduce extra influences and computational cost for classifying the chest X-ray images.
2. Commonly, it is common to find that there are multiple diseases in one image. The relationships among multiple diseases are much more complex compared with the natural images. Thus the discriminative features are not easy to learn for CXR image classification.

Considering the above two critical problems in CXR image classification, we investigate powerful feature learning methods with deep learning techniques in this thesis.

1.1.1 Combining the global and local cues

Many methods generally train a network with global images as input (Wang et al., 2017b; Yao et al., 2017; Guendel et al., 2018). However, thorax disease usually happens in (small) localized areas which are disease-specific. Thus training CNNs using global images may be affected by the (excessive) irrelevant noisy areas. Besides, due to the poor alignment of some CXR images, the existence of irregular borders hinders network performance.

For addressing the above problems, we consider combining the global and local cues to identify thorax diseases and get more precise performance. In Chapter 3, we propose an attention-guided convolutional neural network to classify the CXRs. Specifically, we first learn a global CNN branch using global images. Guided by the attention heatmap generated from the global branch, we inference a mask to crop a discriminative region from the global image. The local region is used for training a local CNN branch. Lastly, we concatenate the last pooling layers of both the global and local branches for fine-tuning the fusion branch. The attention-guided mask inference-based cropping strategy avoids noise and improves alignment in the global branch. AG-CNN also integrates the global cues to compensate for the lost discriminative cues by the local branch. Experiments on the ChestX-ray14 dataset demonstrate that after integrating the local cues with the global information, the average AUC scores are improved by AG-CNN.

1.1.2 Exploiting the correlations of multiple diseases

CXR images are usually labeled with one or more pathologies, which makes the CXR image classification a multi-label problem. Identifying one or more pathologies from a chest X-ray image is often hindered by the pathologies unrelated to the targets. Many works contribute to learn the relationships of multiple diseases in the label space (Yao et al., 2017; Chen et al., 2020a; Guendel et al., 2018). However, considering the problems of sample distribution or imbalance in the label space, it is worth considering if it is appropriate to characterize the relationships of multiple diseases in the feature space.

In chapter 4, a category-wise residual attention learning (CRAL) method is proposed to learn the correlations of multiple diseases in the feature space. CRAL predicts the presence of multiple pathologies in a class-specific attentive view. It aims to suppress the obstacles of irrelevant classes by endowing small weights to the corresponding feature representation. Meanwhile, the relevant features would be strengthened by assigning larger weights.

1.1.3 Learning discriminative features for CXR classification

A robust and stable CXR image analysis system should consider the unique characteristics of CXR images. Particularly, it should be able to: 1) automatically focus on the disease-critical regions (features); 2) adaptively capture the intrinsic relationships among disease features and utilize them to boost the multi-label disease recognition rates jointly.

In chapter 5, we introduce a two-branch architecture, named *ConsultNet*, to achieve those two purposes simultaneously. *ConsultNet* consists of two components. First, an information bottleneck constrained feature selector extracts critical disease-specific features according to the feature importance. Second, a spatial-and-channel encoding-based feature integrator enhances the latent semantic dependencies in the feature space. *ConsultNet* fuses these discriminative features to improve the performance of thorax disease classification in CXRs. Experiments conducted on the ChestX-ray14 and CheXpert datasets demonstrate the effectiveness of the proposed method.

1.2 Thesis Organization

This thesis is organized into the following chapters:

- *Chapter 2* presents a survey of various methods for chest X-ray image classification, particularly emphasizing the lesion area related methods and multi-label learning methods.

- *Chapter 3* proposes to combine the global and local cues to classify the chest X-ray images. An attention-guided mask inference-based strategy is introduced to localize the local discriminative regions. A three-branch neural network is proposed to fuse the global and local features for CXR classification. Experiments on the localization and classification tasks demonstrate the effectiveness of the proposed method.

- *Chapter 4* considers the problem of multi-label chest X-ray image classification. A category-wise residual attention learning method is proposed to learn the correlations of multiple diseases in the feature space. It is expected to strengthen the relevant features and hinder the irrelevant features at the same time. We evaluate the effectiveness of the category-wise residual attention learning method through extensive experiments.

- *Chapter 5* presents a discriminative feature learning method that focuses on the disease-specific and the long-range dependency features. We propose to learn disease-specific features with a variational selective information bottleneck constraint and capture the long-range dependency with a spatial-and-channel encoding method. We achieve the above purposes by integrating them into a ConsultNet. Comprehensive experiments on the chest X-ray image dataset illustrate the effectiveness of ConsultNet.

- *Chapter 6* gives a brief summary of the thesis contents and its contributions. Recommendation for future works is given as well.

Chapter 2

Literature Review

2.1 Chest X-ray image classification with deep learning

Many research works focus on classifying chest X-ray images with deep learning technologies. Wang *et al.* (Wang et al., 2017b) evaluate the performance of several popular deep convolutional neural networks (VGGNet (Simonyan and Zisserman, 2014), GoogLeNet (Szegedy et al., 2015), ResNet (He et al., 2016)) on the ChestX-ray14 dataset. CheXNet (Rajpurkar et al., 2017) modifies the fully connected layer of DenseNet (Huang et al., 2017) and finetunes the parameters of network on the CheXpert (Irvin et al., 2019) dataset. Shen *et al.* (Shen and Gao, 2018) take advantage of the convolutional neural network and the routing-by mechanism in capsule networks to recognize thorax diseases in chest X-ray images. Yan *et al.* (Yan et al., 2018) introduce the squeeze and excitation modules (Hu et al., 2018), multi-map transfer, and max-min pooling techniques into the DenseNet (Huang et al., 2017) framework to improve the chest X-ray image classification performance. Gong *et al.* (Gong et al., 2021) revisit the Gabor filters and propose a deformable Gabor convolution to expand conventional deep networks interpretability and enable complex spatial variations for biomedical image classification. Zhao *et al.* (Zhao et al., 2020) propose to consider the structural attribute of chest X-ray images with cross-chest graphs. Cross-chest graph models the intra-image relationships between different anatomical areas by leveraging the structural information to simulate the doctor's habit of observing different areas. Wang *et al.* (Wang et al., 2020) propose a knowledge-guided deep zoom neural network to leverage prior medical knowledge as training guidance. Chen *et al.* (Chen et al., 2020c) propose a deep hierarchical multi-label method to classify the chest X-ray images.

To accurately recognize the diseases in chest X-ray images, two main popular aspects that researchers focus on are locating the lesion areas and exploring the relationships among multiple diseases, respectively. On the one hand, the diseases are labeled with the pathologies existing in the images, which are highly related to the lesion areas. Recognizing the diseases could benefit from locating the positions of lesion areas. On the other hand, due to the naturally existing relationships among multiple thorax diseases in chest X-ray images, exploring the correlations of these diseases is an intuitive idea to improve the performance of disease classification. In the following, we introduce the chest X-ray images classification method with deep learning from the above two aspects.

2.1.1 Lesion areas related methods

Chest X-ray images are annotated with the diseases existing in the images. The annotations could be considered as semantic labels in medicine. Therefore, a large amount of research works directly concentrate on the feature learning of local lesion areas.

Many studies try to segment or localize the lung/heart areas before classifying chest X-ray images because thorax diseases mainly happen in these regions (Guendel et al., 2018; Liu et al., 2019; Chen et al., 2020b; Liang et al., 2019). For example, Guendel *et al.* (Guendel et al., 2018) classify the thorax diseases with the help of spatial knowledge provided by the PLCO dataset (Gohagan et al., 2000). The spatial knowledge of lesion includes the information about the side (left lung or right lung) and the finer position in each lung (divided into equal fifth). The detailed position information usually does not available in other larger-scale datasets (*e.g.*, ChestX-ray14 (Wang et al., 2017b), CheXpert (Irvin et al., 2019)). Some works turn to learn a lung segmentation model to obtain the positions of lung/heart. Liu *et al.* (Liu et al., 2019) and Chen *et al.* (Chen et al., 2020b) finetune a U-Net (Olaf and Fisher, 2015) variant with the segmentation masks provided by JSRT dataset (Shiraishi et al., 2000; Van Ginneken et al., 2006), and then localize the centered lung areas for further feature learning. Liang *et al.* (Liang et al., 2019) introduce the concept of relative location for thorax disease identification. In this work, we propose a global-local strategy for chest X-ray image classification.

In addition, except for the global information in the image, it is critical to strengthen the local discriminative features due to the diversity of the lesion area in scale and size. Yao *et al.* (Yao et al., 2018) learn from multiple resolution feature maps and generate saliency maps with weak supervision to localize abnormalities of different sizes. Cai *et al.* (Cai et al., 2018) propose an attention mining (AM) strategy to improve the model's sensitivity or saliency to disease patterns. It is motivated by once the most salient disease area is blocked or hidden from the CNN model, AM could pay attention to alternative image regions while still attempting to make correct predictions. Kim *et al.* (Kim et al., 2020) propose an Attend-and-Compare Module (ACM) to capture the differences between an object of interest and its corresponding context, which is helpful for disease localization tasks. Seibold *et al.* (Seibold1 et al., 2020) propose a self-guided loss function (SGL) to improve the accuracy of localization. SGL is used to train a convolutional neural network by increasing the localization confidence and assisting the overall disease identification. Chen *et al.* (Chen et al., 2020a) propose to leverage the region-based and channel-based attention to localize the discriminative features of lesion location and find high weights to the attractive channels. Both the region-based and channel-based attention could focus on the disease-related regions. Hermoza *et al.* (Hermoza et al., 2020) propose to combine the region proposals and saliency detection method to improve the chest X-ray image classification performance. In this work, we first propose a global-local strategy for chest X-ray image classification in Chapter 3, which focus on the local discriminative regions and global information simultaneity.

2.1.2 Multi-label learning

The high diversity of pathologies in the lung/heart area results in that thorax diseases are difficult to diagnose. There are also many works that consider this factor and treat the chest X-ray image classification as a multi-label classification problem. Yao *et al.* (Yao et al., 2017) extract the high-level features with a convolution neural network and encode the multi-label dependency with a Long Short-Term Memory (LSTM)(Hochreiter and Schmidhuber, 1997). Similar with (Yao et al., 2017), Hu *et al.* (Hu et al., 2020) generate the disease predicted sequence with a recurrent neural network aiming to explore the semantic and co-occurrence dependencies among multiple diseases. The complex relationships among the diseases are difficult to represent by the sequential outputs of LSTM or RNN. To overcome this problem, Hu *et al.* (Hu et al., 2020) introduce a historical information module to consider all the generated labels when generating a new label.

Chen *et al.* (Chen et al., 2020a) assume that the relationships among the multiple diseases could be represented by the graph. The nodes of the graph are disease categories, and the edge between two nodes represents the relation between them. A graph convolutional network-based method is leveraged to learn the relationships of diseases (weights of edges). Chen *et al.* (Chen et al., 2020c) propose a multi-label classification method via constructing the hierarchical structure in label space to improve the recognition accuracy. Kumar *et al.* (Kumar et al., 2018) exploit the effect of different loss functions on the chest X-ray image classification and propose a cascaded classifier chain for multi-label disease classification. Most of these methods handle the diseases independently of each other. However, with the situation of one specific pathology with a small number of samples, learning with existing loss functions would lose important discriminative information for that pathology. Therefore, Mo *et al.* (Mo and Cai, 2019) propose a weighted entropy loss function to learn the correlations among the labels by making full use of small amount samples. In contrast with these existing methods, we focus on learn the correlations of multiple diseases in the feature space in Chapter 4.

In Chapter 5, we further investigate the chest X-ray image classification problem from the view of discriminative feature learning. A two-branch convolutional neural network, ConsultNet, is proposed to automatically filter the disease-specific features and encode the long-range dependencies of features. ConsultNet consists of a varialtional selective information bottleneck (VSIB) branch and a spatial-and-channel encoding branch. VSIB introduces a spatial-wise and channel-wise based attention mechanism into the Variational Information Bottleneck (VIB) to enforce the network to select critical, disease-specific features for chest X-ray image classification. For the view of deep networks, the goal of information bottleneck (TISHBY, 1999) (IB) is to learn an encoding by maximizing the mutual information between the latent representation $\boldsymbol{Z}$ of input $\boldsymbol{X}$ and the class $\boldsymbol{Y}$. A natural constraint to apply is on the mutual information between the input features $\boldsymbol{X}$ and the latent representation $\boldsymbol{Z}$, $I(\boldsymbol{X}, \boldsymbol{Z}) \leq I_c$, where I_c is the information constraint. Training a deep convolutional neural network minimizes this function, and thus can obtain a maximally compact latent representation $\boldsymbol{Z}$ that is informative about the class $\boldsymbol{Y}$. That is, the latent representation $\boldsymbol{Z}$ is expected to contain as much information as the class $\boldsymbol{Y}$ ($I(\boldsymbol{Y}, \boldsymbol{Z})$ is large), but not tell more about $\boldsymbol{X}$ that is necessary to correctly estimate $\boldsymbol{Y}$ ($I(\boldsymbol{X}, \boldsymbol{Z})$ is small). This is equivalent to upperbound a Kullback-Leibler (KL) divergence between the joint probability $P(\boldsymbol{X}; \boldsymbol{Z})$ and the product of the marginals $P(\boldsymbol{X}) \cdot P(\boldsymbol{Z})$ to a specific bottleneck value I_c. Part of our work is related to the recently proposed InfoMask pneumonia localization method (Taghanaki and Havaei, 2019). Unlike (Taghanaki and Havaei, 2019), which aims at the lesion area localization task, we mainly focus on disease classification in a multi-label learning framework. Technically, (Taghanaki and Havaei, 2019) designs a mask layer to detect the positions of pneumonia by introducing the vanilla VIB principle. In comparison, VSIB introduces a selective mechanism into the vanilla VIB principle to learn compact, disease-specific features. The proposed selective mechanism is achieved by considering the spatial-wise and channel-wise attention and used to measure the feature importance. Due to the implicit learning for disease-related features in VSIB, we do not enforce the model to detect the positions of the whole lesion area while only focus on the most discriminative regions (usually a small part of the lesion area) for classification.

Chapter 3

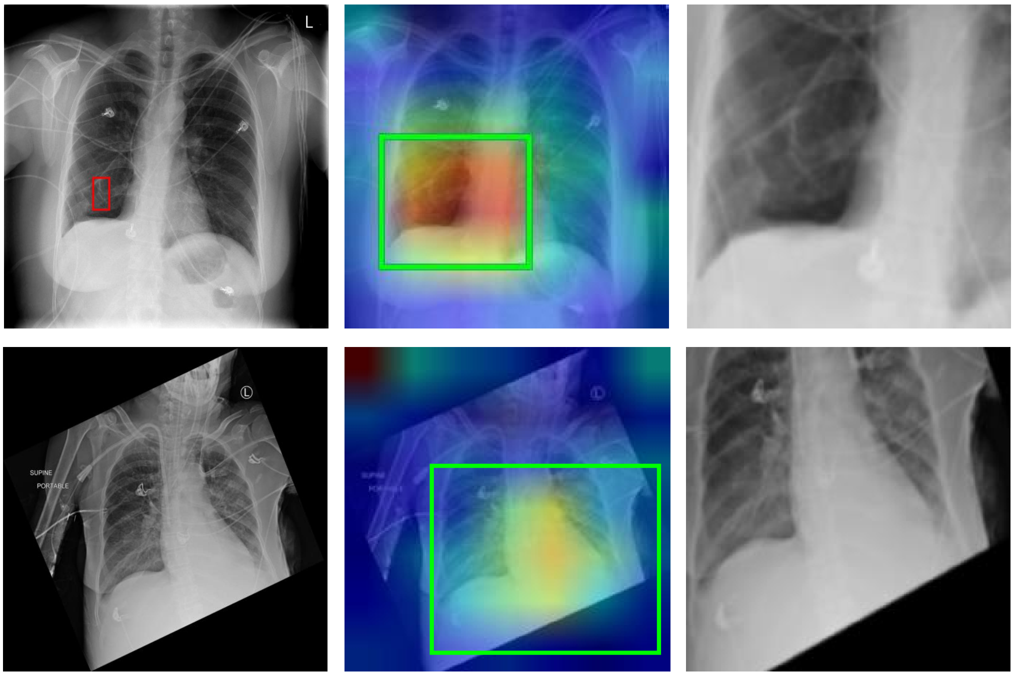
Attention Guided Convolution Neural network

In this chapter, we consider the the task of thorax disease diagnosis on chest X-ray (CXR) images by combining the global and local cues with deep learning technologies. An attention guided convolution neural network (AG-CNN) is proposed to integrate the global and local cues and identify thorax diseases. In this chapter, we first introduce the motivation of this work. And then we describe the proposed AG-CNN framework. Finally, we demonstrate the effectiveness of AG-CNN on one large scale chest X-ray dataset.

3.1 Motivation

Thorax disease usually happens in (small) localized areas which are disease specific. Thus training CNNs using global image may be affected by the (excessive) irrelevant noisy areas. Besides, due to the poor alignment of some CXR images, the existence of irregular borders hinders the network performance.

Several existing works on CXR classification typically employ the *global image* for training (Wang et al., 2017b; Yao et al., 2017; Kumar et al., 2018). However, the global learning strategy can be compromised by two problems. On the one hand, using the global image for classification may include a considerable level of noise outside the lesion area. As shown in Fig. 3.1 (the first row), the lesion area can be very small (red bounding box) compared with the global image. These large numbers of healthy regions make the deep networks hard to focus on the local lesion area, and the positions of disease regions are also unpredictable. This problem is rather different from generic image classification (Deng et al., 2009), where the object of interest is usually positioned in the image center.



(a) original global image (b) heat map (c) cropped local image

Figure 3.1 : Two images from the ChestX-ray14 dataset. (a) The global images. (b) heatmaps extracted from a specific convolutional layer. (c) The cropped images from (a) guided by (b).

Besides, due to the large inter-class similarity of chest X-ray images, it is hard for the deep networks to capture the subtle discrepancies of different classes in the whole images, especially when the critical lesion areas are very small. Considering this fact, it is beneficial to induce the network to focus on the lesion regions when making predictions. On the other hand, due to the variations of capturing condition, *e.g.*, the posture of the patient, or the small size of the child body, the CXR images may undergo distortion or misalignment. Fig. 3.1 (the second row) presents a misalignment example. This human body is relatively small, and a large number of regions are all black in the image. The irregular image borders may exist a non-negligible effect on classification accuracy. In real scenarios, some chest X-ray images could not be re-captured. Thus, the computer-aided diagnosis system is expected to make accurate predictions on the existing images. That is, the diagnosis algorithm should be robust to the quality of the chest X-ray images. Therefore, it is desirable to discover the salient lesion regions and thus alleviate the impact of such misalignment.

In this chapter, we consider both the original global image and the cropped local image for classification, so that 1) the noise contained in non-lesion area is less influencing, and 2) the misalignment can be reduced. Though there is a high activation region in the top-left corner of heatmap (second row), the proposed maximum connected region cropping strategy could ensure to avoid selecting such obvious noisy region.

To address the problems caused by merely relying on the global CXR image, this chapter introduces a three-branch attention guided convolutional neural network (AG-CNN) which integrates the global and local cues to classify the lung or heart diseases. AG-CNN is featured in two aspects. First, it has a focus on the local lesion regions which are disease specific. Generally, such a strategy is particularly effective for diseases such as ”Nodule”, which has a small lesion region. In this manner, the impact of the noise in non-disease regions and misalignment can be alleviated. Second, AG-CNN has three branches, *i.e.*a global branch, a local branch and a fusion branch. While the local branch exhibits the attention mechanism, it may lead to information loss in cases where the lesion areas are distributed in the whole images, such as Pneumonia. Therefore, a global branch is needed to compensate for this error. We show that the global and local branches are complementary to each other and, once fused, yield favorable accuracy to the state of the art.

The working mechanism of AG-CNN is similar to that of a radiologist. We first learn a global branch that takes the global image as input: a radiologist may first browse the whole CXR image. Then, we discover and crop a local lesion region and train a local branch: a radiologist will concentrate on the local lesion area after the overall browse. Finally, the global and local branches are fused to fine-tune the whole network: a radiologist will comprehensively consider the global and local information before making decisions.

The main contributions are summarized as follows.

- Chest X-ray images classification suffers from exploring the distinct lesion areas. A visual attention-guided region inference approach is proposed to localize the local lesion area. The attention-guided method crops the discriminative regions to classify the chest X-ray image and thus corrects the image alignment and reduces the impact of noise.
- An attention-guided convolutional neural network is proposed to diagnose thorax diseases. AG-CNN simulates the human expert in terms of attention. The latter not only focuses on the global appearance but also looks for the specific lesion areas, before combining the two perspectives to reach a final decision. AG-CNN employs and fuses global and local information to mimic the human diagnosing procedure and reporting competitive accuracy.

3.2 Method

In this section, we describe the framework of AG-CNN.

3.2.1 Structure of AG-CNN

The architecture of AG-CNN is presented in Fig. 3.2. Basically, it has two major branches, *i.e.*the global and local branches, and a fusion branch. Both the global and local branches are classification networks that predict whether the pathologies are present or not. Given an image, the global branch is first fine-tuned from a classification CNN using the global image. Then, we crop an attractive region from the global image and train it for classification on the local branch. Finally, the last pooling layers of both the global and local branches are concatenated for fine-tuning the fusion branch.

Global and local branches. The global branch informs the underlying CXR infor-

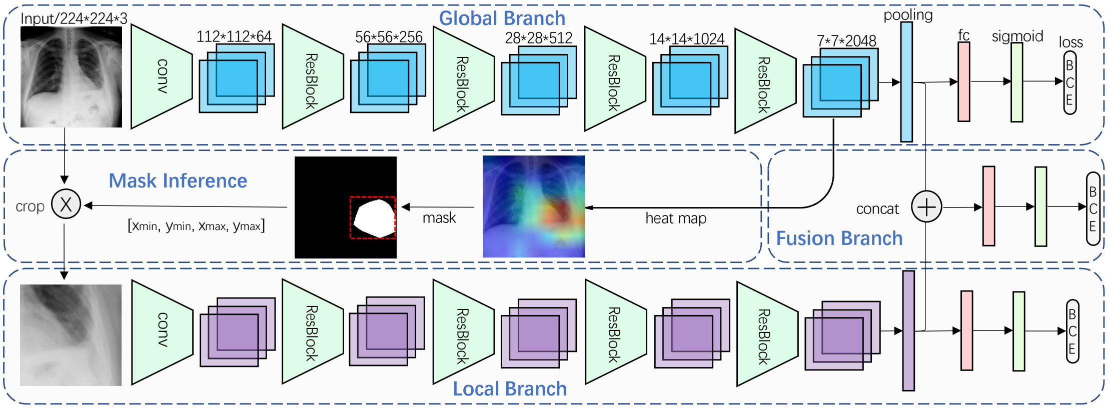


Figure 3.2 : Overall framework of the attention guided convolutional neural network (AG-CNN, showing ResNet-50 as backbone). ”BCE” represents binary cross entropy loss. The spatial resolution of heatmap generated from the last convolutional layer of the global branch is 7×7. Then we resize the heatmap to 224×224 by bilinear interpolation. The input image is added to the heatmap for visualization.

mation derived from the global image as input. In the global branch, we train a variant of ResNet-50 (He et al., 2016) as the backbone model. It consists of five down-sampling blocks, followed by a global max pooling layer and a C-dimensional fully connected (FC) layer for classification. At last, a sigmoid layer is added to normalize the output vector $p_g(c|I)$ of FC layer by

$$\widetilde{p_g}(c|I) = 1/(1 + exp(-p_g(c|I))), \tag{3.1}$$

where I is the global image. $\widetilde{p_g}(c|I)$ represents the probability score of I belonging to the c^{th} class, $c \in \{1, 2, ..., C\}$. We optimize the parameter W_g of global branch by minimizing the binary cross-entropy (BCE) loss:

$$\mathcal{L}(W_g) = -\frac{1}{C}\sum_{c=1}^{C} l_c log(\widetilde{p_g}(c|I)) + (1 - l_c)log(1 - \widetilde{p_g}(c|I)), \tag{3.2}$$

where l_c is the groundtruth label of the c^{th} class, C is the number of pathologies.

On the other hand, the local branch focuses on the lesion area and is expected to alleviate the drawbacks of only using the global image. In more details, the local branch possesses the same convolutional network structure with the global branch. Note that, these two branches do not share weights since they have distinct purposes. We denote the probability score of local branch as $\widetilde{p_l}(c|I_c)$, W_l as the parameters of local branch. Here, I_c is the input image of local branch. We perform the same normalization and optimization as the global branch.

Fusion branch. The fusion branch first concatenates the Pool5 outputs of the global and local branches. The concatenated layer is connected to a 15-dimensional FC layer for final classification. The probability score is $\widetilde{p_f}(c|[I, I_c])$. We denote W_f as the parameters of fusion branch and optimize W_f by Eq. 3.2.

3.2.2 Attention Guided Mask Inference

A binary mask is constructed to locate the discriminative regions for classification in the global image. It is produced by performing thresholding operations on the feature maps, which can be regarded as an attention process. This process is described below.

Given a global image, let $f_g^k(x, y)$ represent the activation of spatial location (x, y) in the k-th channel of the output of the last convolutional layer, where $k \in \{1, ..., K\}$, $K = 2,048$ in ResNet-50. g denotes the global branch. We first take the absolute value of the activation values $f_g^k(x, y)$ at position (x, y). Then the attention heatmap H_g is generated by counting the maximum values along channels,

$$H_g(x, y) = \max_k(|f_g^k(x, y)|), k \in \{1, ..., K\}. \tag{3.3}$$

The values in H_g directly indicate the importance of the activations for classification. In Fig. 3.1(b) and Fig. 3.3 (the second row), some examples of the heatmaps are shown. We observe that the discriminative regions (lesion areas) of the images are activated. heatmap can be constructed by computing different statistical values across the channel dimensions, such as L1 distance $\frac{1}{K}\sum_{k=1}^{K}|f_g^k(x, y)|$ or L2 distance $\frac{1}{K}\sqrt{\sum_{k=1}^{K}(f_g^k(x, y))^2}$. Different statistics results in subtle numerical differences in heatmap, but may not effect the

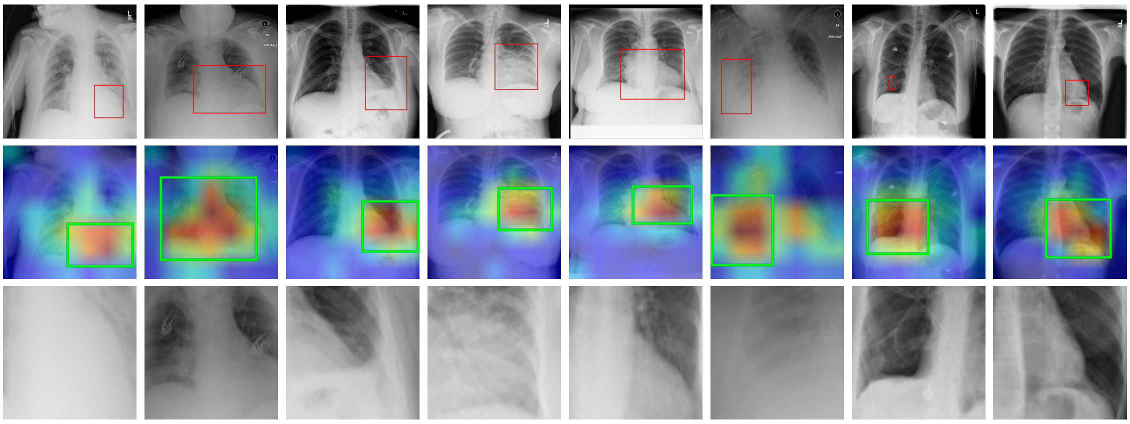


Figure 3.3 : The process of lesion area generation. (**Top:**) global CXR images of various thorax diseases for the global branch. Note that we do not use the bounding boxes for training or testing. (**Middle:**) corresponding visual examples of the output of the mask inference process. Higher/lower response is denoted with red/blue. (**Bottom:**) cropped and resized images from the green bounding boxes which are fed to the local branch.

classification significantly. Therefore, we compute heatmap with Eq. 3.3 in our experiment. The comparison of these statistics is presented in Section 3.3.2.

We design a binary mask M to locate the regions with large activation values. If the value of a certain spatial position (x, y) in the heatmap is larger than a threshold τ , the value at corresponding position in the mask is assigned with 1. Specifically,

$$M(x,y) = \begin{cases} 1, & H_g(x,y) > \tau \\ 0, & \text{otherwise} \end{cases}, \tag{3.4}$$

where τ is the threshold that controls the size of attended region. A larger τ leads to a smaller region, and vice versa. With the mask M, we draw a maximum connected region that covers the discriminative points in M. The maximum connected region is denoted as the minimum and maximum coordinates in horizontal and vertical axis $[x_{min}, y_{min}, x_{max}, y_{max}]$. At last, the local discriminative region I_c is cropped from the input image I and is resized to the same size as I. We visualize the bounding boxes and cropped patches with

Algorithm 1 Attention Guided CNN Procedure

```
Input: Input image $I$; Label vector $L$; Threshold $\tau$.
Output: Probability score $\widetilde{p_f}(c|[I, I_c])$.
Initialization: The global and local branch weights.
Learning $W_g$ with $I$, computing $\widetilde{p_g}(c|I)$, optimizing by Eq. 3.2 (Stage I);
Computing mask $M$ and the bounding box coordinates $[x_{min}, y_{min}, x_{max}, y_{max}]$, cropping
  out $I_c$ from $I$;
Learning $W_l$ with $I_c$, computing $\widetilde{p_l}(c|I_c)$, optimizing by Eq. 3.2 (Stage II);
Concentrating $Pool_g$ and $Pool_l$, learning $W_f$, computing $\widetilde{p_f}(c|[I, I_c])$, optimizing by
  Eq. 3.2.
```

$\tau = 0.7$ in Fig. 3.3. The attention informed mask inference method is able to locate the regions (green bounding boxes) which are reasonably close to the groundtruth (red bounding boxes).

3.2.3 Training Strategy of AG-CNN

A three-stage training scheme is adopt for AG-CNN.

Stage I. Using the global images, we fine-tune the global branch network pretrained by ImageNet. $\widetilde{p_g}(c|I)$ is normalized by Eq. 3.1.

Stage II. Once the local image I_c is obtained by mask inference with threshold τ, we feed it into the local branch for fine-tuning. $\widetilde{p_l}(c|I_c)$ is also normalized by Eq. 3.1. When we fine-tune the local branch, the weights in the global branch are fixed.

Stage III. Let $Pool_g$ and $Pool_l$ represent the Pool5 layer outputs of the global and local branches, respectively. We concatenate them for a final stage of fine-tuning and normalize the probability score $\widetilde{p_f}(c|[I, I_c])$ by Eq. 3.1. Similarly, the weights of previous two branches are fixed when we fine-tune the weights of fusion branch.

In each stage, we use the model with the hyper-parameter τ with the highest AUC score on the validation set for testing. The overall AG-CNN training procedure is presented in Algorithm 1. Variants of training strategy may influence the performance of AG-CNN. We discuss it in Section 3.3.2.

3.3 Experiment

This section evaluates the performance of the proposed AG-CNN. The experimental dataset, evaluation protocol and the experimental settings are introduced first. Section 3.3.2 demonstrates the performance of global and local branches and the effectiveness of fusing them. Furthermore, comparison of AG-CNN and the state of the art is presented in Table. 3.1. In Section. 3.3.3, we analyze the parameter impacts in mask inference.

Dataset. We evaluate the AG-CNN framework using the ChestX-ray14 (Wang et al., 2017b). ChestX-ray14 collects 112,120 frontal-view images of 30,805 unique patients. 51,708 images of them are labeled with up to 14 pathologies, while the others are labeled as "No Finding".

Evaluation protocol. In our experiment, we randomly shuffle the dataset into three subsets: 70% for training, 10% for validation and 20% for testing. Each image is labeled with a 15-dim vector $\mathbf{L} = [l_1, l_2, ..., l_c, ..., l_C]$ in which $l_c \in \{0, 1\}, C = 15$. l_{15} represents the label with "No Finding".

3.3.1 Experimental details

For training (any of the three stages), we perform data augmentation by resizing the original images to 256×256, randomly resized cropping to 224×224, and random horizontal flipping. The ImageNet mean value is subtracted from the image. When using ResNet-50 as backbone, we optimize the network using SGD with a mini-batch size of 126, 64, 64 for global, local and fusion branch, respectively. But for DenseNet-121, the network is optimized with a mini-batch of 64, 32, and 32, respectively. We train each branch for 50 epochs. The learning rate starts from 0.01 and is divided by 10 after 20 epochs. We use a weight decay of 0.0001 and a momentum of 0.9. During validation and testing, we also resize the image to 256×256, and then perform center cropping to obtain an image of size 224×224. Except in Section 3.3.3, we set τ to 0.7 which yields the best performance on the validation set. We implement the proposed framework with Pytorch. We train the network on a computer with NVIDIA TITAN Xp GPUs. The training process of global or local branch takes about 6 hours on the ChestX-ray14 dataset (more than 80,000 training samples).

3.3.2 Evaluation

We evaluate our method on the ChestX-ray14 dataset. Mostly, ResNet-50 (He et al., 2016) is used as backbone, but the AUC and ROC curve obtained by DenseNet-121 (Huang et al., 2017) are also presented.

Global branch (baseline) performance. We first report the performance of the baseline, *i.e.*the global branch. Results are summarized in Table. 3.1 and Fig. 3.9. The average AUC across the 14 thorax diseases arrives at 0.841 and 0.840, using ResNet-50 and DenseNet-121, respectively. For both backbone networks, this is a competitive accuracy compared with the previous state of the art. Except Herina, the AUC scores of the other 13 pathologies are very close to or even higher than (Rajpurkar et al., 2017). Moreover, we observe that Infiltration has the lower recognition accuracy (0.728 and 0.717 for ResNet-50 and DenseNet-121). This is because the diagnosis of Infiltration mainly relies on the texture change among the lung area, which is challenging to recognize. The disease Cardiomegaly achieves higher recognition accuracy (0.904 and 0.912 for ResNet-50 and DenseNet-121, respectively), which is characterized by the relative solid region (heart).

Table 3.1 : Comparison results of various methods on ChestX-ray14.

Method	CNN	Atel	Card	Effu	Infi	Mass	Nodu	Pne1	Pne2	Cons	Edem	Emph	Fibr	PT	Hern	Mean
Wang *et al.*(Wang et al., 2017b)	R-50	0.716	0.807	0.784	0.609	0.706	0.671	0.633	0.806	0.708	0.835	0.815	0.769	0.708	0.767	0.738
Yao *et al.*(Yao et al., 2017)	D-/	0.772	0.904	0.859	0.695	0.792	0.717	0.713	0.841	0.788	0.882	0.829	0.767	0.765	0.914	0.803
Rajpurkar *et al.*(Rajpurkar et al., 2017)	D-121	0.821	0.905	0.883	0.720	0.862	0.777	0.763	0.893	0.794	0.893	0.926	0.804	0.814	0.939	0.842
Kumar *et al.*(Kumar et al., 2018)	D-161	0.762	0.913	0.864	0.692	0.750	0.666	0.715	0.859	0.784	0.888	0.898	0.756	0.774	0.802	0.795
Global branch (baseline)	R-50	0.818	0.904	0.881	0.728	0.863	0.780	0.783	0.897	0.807	0.892	0.918	0.815	0.800	0.889	0.841
Local branch	R-50	0.798	0.881	0.862	0.707	0.826	0.736	0.716	0.872	0.805	0.874	0.898	0.808	0.770	0.887	0.817
AG-CNN	R-50	0.844	0.937	0.904	0.753	0.893	0.827	0.776	0.919	0.842	0.919	0.941	0.857	0.836	0.903	0.868
Global branch (baseline)	D-121	0.832	0.906	0.887	0.717	0.870	0.791	0.732	0.891	0.808	0.905	0.912	0.823	0.802	0.883	0.840
Local branch	D-121	0.797	0.865	0.851	0.704	0.829	0.733	0.710	0.850	0.802	0.882	0.874	0.801	0.769	0.872	0.810
AG-CNN	D-121	0.853	0.939	0.903	0.754	0.902	0.828	0.774	0.921	0.842	0.924	0.932	0.864	0.837	0.921	0.871

[*] Each pathology is denoted with its first four characteristics, *e.g.*. Pneumonia and Pneumothorax are denoted as *Pneu1* and *Pneu2*, respectively. PT represents Pleural Thickening. We report the performance with parameter $\tau = 0.7$. For each column, the best and second best results are highlighted in red and blue, respectively.

Performance of the local branch. We crop the most discriminative region to improve the classification accuracy. The local branch is trained on the cropped and resized discrimative patches, which is supposed to provide attention mechanisms complementary to the global branch. The performance of the local branch is demonstrated in Table. 3.1 and Fig. 3.9.

Using ResNet-50 and DenseNet-121, the average AUC score is 0.817 and 0.810, respectively, which is higher than (Wang et al., 2017b; Kumar et al., 2018). Despite of being competitive, the local branch yields lower accuracy than the global branch. The probable reason for this observation is that the lesion region estimation and cropping process may lead to information loss which is critical for recognition. So the local branch may suffer from inaccurate estimation of the attention area. Generally, the area where the lung is inflamed is relative large and its corresponding attention heatmap shows a scattered distribution. With a higher value of τ, only a very small patch is cropped in original image. For the classes “Hernia” and “Consolidation”, the local and global branch yield very similar accuracy. We speculate that the cropped local patch is consist with the lesion area in the global image.

To illustrate the effectiveness of the cropping strategy of AG-CNN, we test the localization accuracy using the ground truth bounding boxes provided by (Wang et al., 2017b). Intersection over Union (IoU) is computed between the cropped region in AG-CNN and the ground truth. A correct localization result is defined by requiring IoU $>$ T(IoU), where T(IoU) is a threshold. We measure the effect of the parameter τ and T(IoU) in AG-CNN. Fig. 3.4 presents the localization accuracy of different τ in $\{0.2, 0.3, ..., 0.9\}$. In each sub-figure, different color represents the threshold of IoU when measuring the accuracy of the predicted bounding box. As shown in Fig. 3.4, lower τ produces worse localization accuracy. And at the same time, when T(IoU) becomes larger than 0.3, the localization accuracy of most pathologies reduces to zero. In general localization task, the T(IoU) is expected at least greater than 0.5. Therefore, we expect that the selected τ could provide a

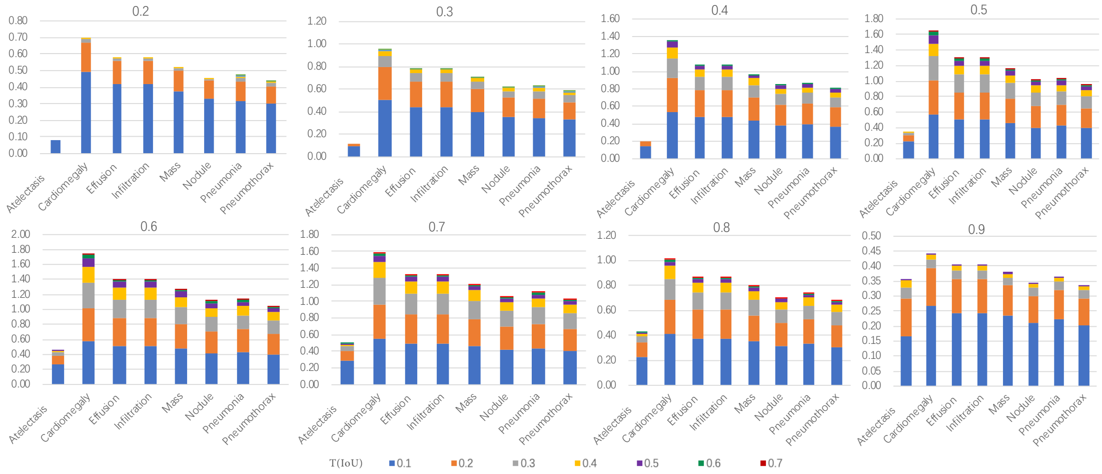


Figure 3.4 : The localization accuracy of different threshold of τ. Each sub-figure is the accuracy for different τ. And in each sub-figure, different color represents the threshold of IoU (T(IoU)) when measuring the accuracy of the predicted bounding box. Better view as zoomed.

relatively larger localization accuracy to satisfy the localized region near to the true lesion area. When τ in $\{0.5, 0.6, 0.7\}$, the localization accuracy are better than others. While τ is larger than 0.8, the accuracy drops significantly. Thus, $\{0.5, 0.6, 0.7\}$ is suggested for the hyperparameter τ.

We compare the localization accuracy with existing methods and the results are summarized in Table. 3.2. Under different IoU thresholds, the localization accuracy of our method is consistently higher than (Wang et al., 2017b). Because both our method and (Wang et al., 2017b) only use image-level labels, this comparison could be regarded as fair. Compared with (Li et al., 2018), our method is inferior. The reason is that (Li et al., 2018) uses additional ground truth bounding boxes which we do not use. Therefore, it is expected that (Li et al., 2018) has a higher localization accuracy due to its usage of additional supervision. However, we also notice that our method is advantageous in localizing the small lesions for the disease “Nodule”: the accuracy of “Nodule” lesion region localization significantly exceeds (Li et al., 2018) under all the thresholds. Besides, the

Table 3.2 : Comparison of localization accuracy.

T(IoU)	Model	Atel	Card	Effu	Infi	Mass	Nodu	Pne1	Pne2	mean
0.1	(Wang et al., 2017b)	0.69	0.94	0.66	0.71	0.40	0.14	**0.63**	0.38	0.57
	(Li et al., 2018)	**0.71**	**0.98**	**0.87**	**0.92**	**0.71**	0.40	0.60	**0.63**	0.73
	Ours	0.48	0.71	0.67	0.67	**0.65**	**0.58**	0.62	0.58	0.62
0.2	(Wang et al., 2017b)	0.47	0.68	0.45	0.48	0.26	0.05	0.35	0.23	0.37
	(Li et al., 2018)	**0.53**	**0.97**	**0.76**	**0.83**	**0.59**	0.29	**0.50**	**0.51**	0.62
	Ours	0.27	0.59	0.50	0.50	0.48	**0.42**	0.45	0.41	0.45
0.3	(Wang et al., 2017b)	0.24	0.46	0.30	0.28	0.15	0.04	0.17	0.13	0.22
	(Li et al., 2018)	**0.36**	**0.94**	**0.56**	**0.66**	**0.45**	0.17	**0.39**	**0.44**	0.50
	Ours	0.14	0.50	0.41	0.41	0.37	**0.33**	0.34	0.32	0.35
0.4	(Wang et al., 2017b)	0.09	0.28	0.20	0.12	0.07	0.01	0.08	0.07	0.12
	(Li et al., 2018)	**0.25**	**0.88**	**0.37**	**0.50**	**0.33**	0.11	**0.26**	**0.29**	0.37
	Ours	0.06	0.39	0.30	0.30	0.27	**0.24**	0.25	0.23	0.25
0.5	(Wang et al., 2017b)	0.05	0.18	0.11	0.07	0.01	0.01	0.03	0.03	0.06
	(Li et al., 2018)	**0.14**	**0.84**	**0.22**	**0.30**	**0.22**	0.07	**0.17**	**0.19**	0.27
	Ours	0.03	0.21	0.16	0.16	0.14	**0.13**	0.14	0.12	0.14
0.6	(Wang et al., 2017b)	0.02	0.08	0.05	0.02	0.00	0.01	0.02	0.03	0.03
	(Li et al., 2018)	**0.07**	**0.73**	**0.15**	**0.18**	**0.16**	0.03	**0.10**	**0.12**	0.19
	Ours	0.00	0.09	0.06	0.06	0.06	**0.05**	0.06	0.06	0.06
0.7	(Wang et al., 2017b)	0.01	0.03	0.02	0.00	0.00	0.00	0.01	0.02	0.01
	(Li et al., 2018)	**0.04**	**0.52**	**0.07**	**0.09**	**0.11**	**0.01**	**0.05**	**0.05**	0.12
	Ours	0.00	0.02	0.02	0.02	0.01	**0.01**	0.01	0.01	0.01

[*] Note that (Wang et al., 2017b) and ours are supervised by image-level labels, while (Li et al., 2018) is supervised by both image-level labels and partially bounding box-level annotations.

performance of some pathologies, such as “Mass”, “Pneumonia”, and “Pneumothorax” are very close to (Li et al., 2018). But we also notice that the performance of “Atelectasis” is inferior to (Wang et al., 2017b). And for “Cardiomegaly”, the localization accuracy is lower than (Wang et al., 2017b) when T(IoU) is less than 0.3, while it is slightly higher than (Wang et al., 2017b) when T(IoU) is greater or equal to 0.3. We analyze that the main reason may be the AG-CNN focuses on the small discriminative regions for classification while not the whole region of interests. Therefore, the cases of “Atelectasis” and “Cardiomegaly” could happen when the features learned by AG-CNN cover parts of the whole lesion area. Overall speaking, in terms of disease localization, our method yields higher accuracy compared with (Wang et al., 2017b) under the same setting, which serves as an explanation of our superior performance.

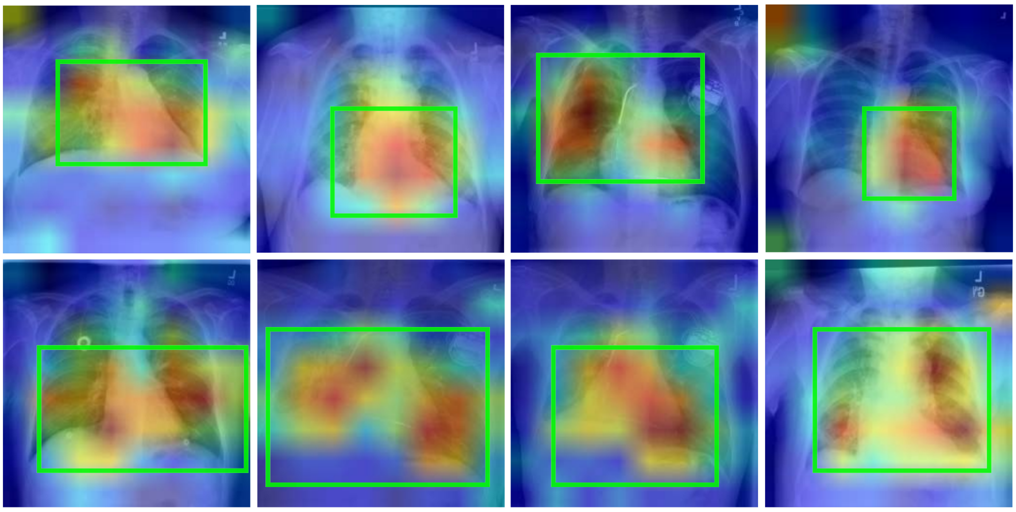


Figure 3.5 : Examples of heatmaps for “no finding” images. The cropped regions are denoted by green bounding boxes.

For the “no finding” images, AG-CNN can also learn the corresponding masks. The automatically discovered ROIs in the “no finding” class contain discriminative information of this class. These ROIs filter out some noisy and misaligned regions and force the network to focus on these important regions during recognition. Thus, the ROIs help to

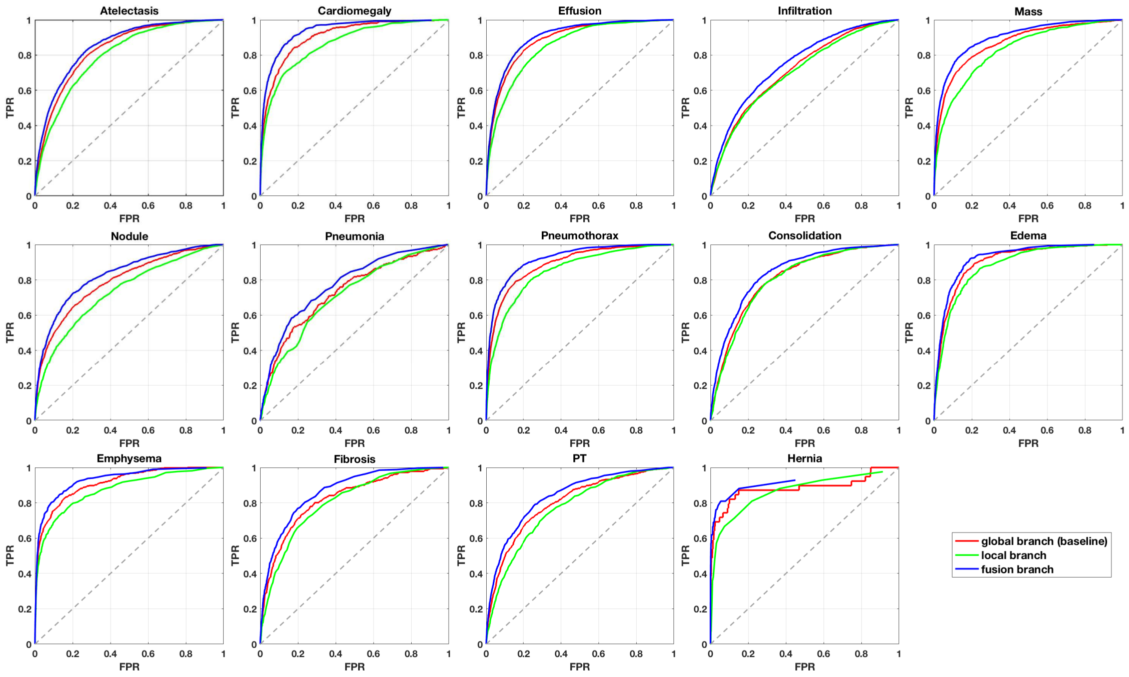


Figure 3.6 : ROC curves of the global, local and fusion branches (DenseNet-121 as backbone) over the 14 pathologies. The corresponding AUC values are given in Table. 3.1. We observe that fusing global and local information yields clear improvement.

distinguish “no finding” from the other 14 pathologies. The “no finding” class plays a role like the background class in object detection. We visualize some cropped region on the heatmap in Fig. 3.5.

Discussions of the proposed cropping strategy. The attention maps generated from the global branch are used to guide the input of the local branch. Once the cropped regions fail to locate the accurate lesion areas, it would directly reduce the recognition accuracy of the local branch. We visualize some heatmaps of the global branch and the corresponding cropped regions in Fig. 3.7 and discuss the cropping strategy compared with the CAM or Grad-CAM. The red and green bounding boxes represent the ground truths and the cropped regions, respectively.

In AG-CNN, we ideally expect the proposed cropping strategy could crop the regions that cover the lesion areas but not accurately locate these lesion areas, as shown in the

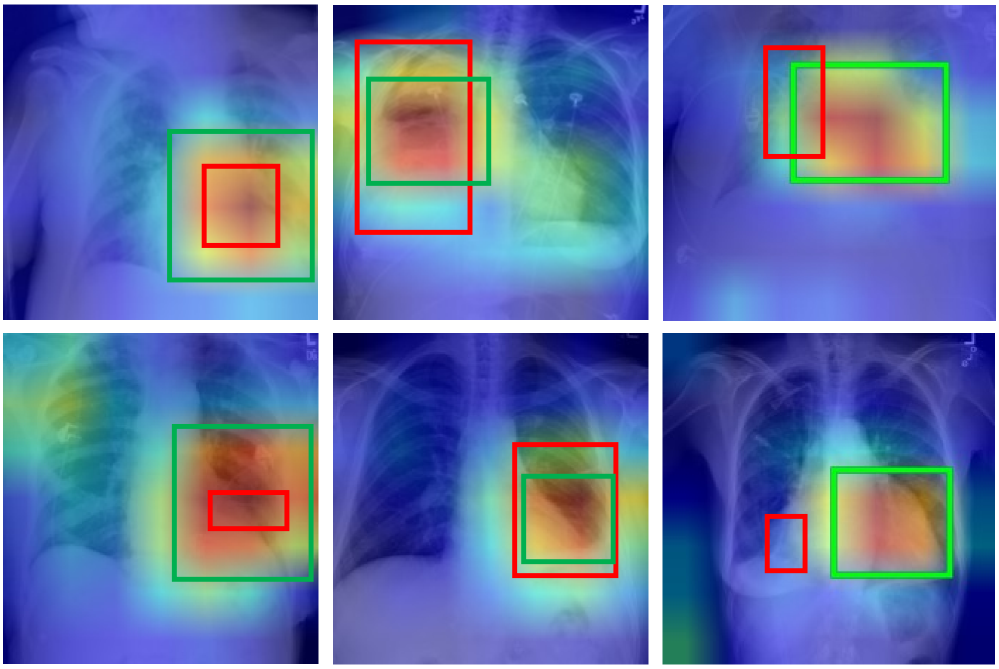


Figure 3.7 : The visualized cropped regions and the lesion areas. The red bounding boxes are the ground truths of lesion areas, and the green bounding boxes are the cropped regions in AG-CNN.

first column of Fig. 3.7. Both the global and the local branches could produce accurate predictions in this situation. However, the global branch or the cropping strategy might not always satisfy such requirements. For example, in the second column of Fig. 3.7, the cropped regions miss some information of the lesion areas. Fortunately, the global branch of AG-CNN could compensate for the missing information for final prediction in the fusion branch, even if the cropped regions fail to locate the accurate lesion areas. The worst case is that the cropped regions fail to locate the lesion areas, as shown in Fig. 3.7 (the last column).

It is well known that CAM or Grad-CAM could localize the object regions based on the prediction scores, which is more precise than the proposed attention-guided cropping strategies. While one problem using CAM or Grad-CAM is that once the cropped regions miss the lesion areas, the global branch might not compensate for the losing information, either. That is also a consideration that not use the CAM or Grad-CAM to extract the feature maps in the local branch. The proposed attention-guided cropping strategy aims to locate the high response regions in the global branch and not compulsorily request to localize accurate lesion areas. Such relaxation could tolerate some influences of failure cases to the whole framework.

Effectiveness of fusing global and local branches. We illustrate the effectiveness of the fusion branch, which yields the final classification results of our model. The observations are consistent across different categories and the two backbones. We present the ROC curves of 14 pathologies with these two backbones in Fig. 3.8. For both ResNet-50 and DenseNet-121, the fusion branch, *i.e.*, AG-CNN, outperforms both the global branch and local branch. For example, when using ResNet-50, the performance gap from AG-CNN to the global and local branches is 0.027 and 0.051, respectively. Specifically AG-CNN (with DenseNet-121 as backbone) surpasses the global and local branches for all 14 pathologies. Fig. 3.6 presents the ROCs of three branches of each pathology in ChestX-ray14.

We conduct another experiment, inputting a global image into both the global and local branches to verify the effectiveness of fusing global and local cues. The same experimental settings with Section 3.3.1 are performed. Three branches are trained together with ResNet-50 as backbone. The average AUC of global, local and fusion branches achieve to 0.845, 0.846 and 0.851, respectively. The AUC is lower 0.017 compared with inputting a local patch into the local branch. The results show that AG-CNN is superior than both global and local branches. In particular, the improvement is benefit from the local discriminative region instead of increasing the number of network parameters.

Comparison with the state of the art. We compare our results with the state-of-the-art methods (Wang et al., 2017b; Yao et al., 2017; Kumar et al., 2018; Rajpurkar et al., 2017) on the ChestX-ray14 dataset. Wang *et al.*(Wang et al., 2017b) classify and

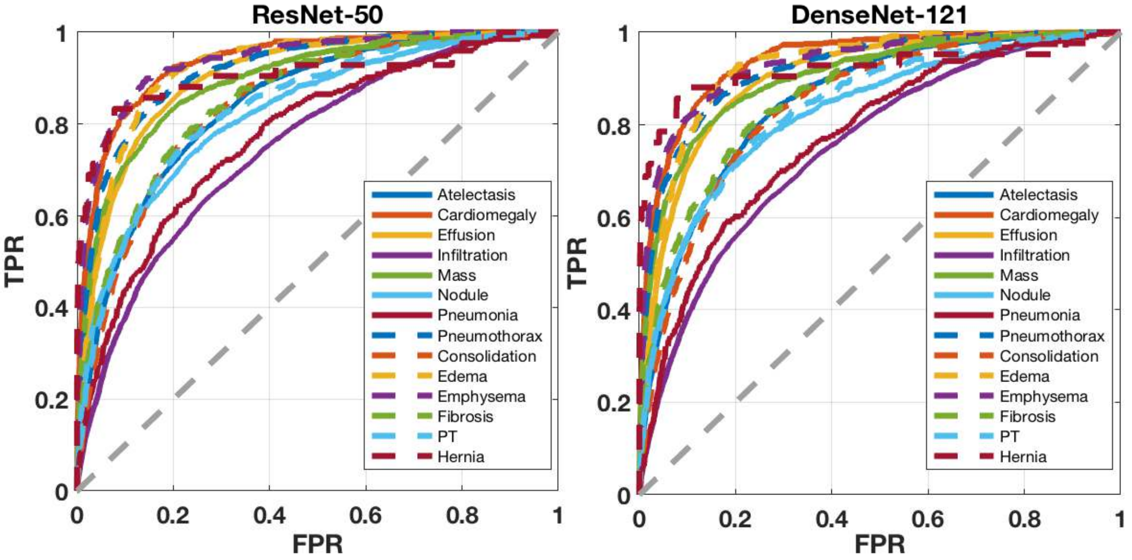


Figure 3.8 : ROC curves of AG-CNN on the 14 diseases (ResNet-50 and DenseNet-121 as backbones, respectively).

localize the thorax disease in a unified weakly supervised framework. The reported results from Yao *et al.*(Yao et al., 2017) are based on the model in which labels are considered independent. Kumar *et al.*(Kumar et al., 2018) try different boosting methods and cascade the previous classification results for multi-label classification.

Comparing with these methods, **this paper contributes new state of the art to the community: average AUC = 0.871.** AG-CNN exceeds the previous state of the art (Rajpurkar et al., 2017) by 2.9%. AUC scores of pathologies such as *Cardiomegaly* and *Inftration* are higher than (Rajpurkar et al., 2017) by about 0.03. AUC scores of *Mass*, *Fibrosis* and *Consolidation* surpass (Rajpurkar et al., 2017) by about 0.05. Furthermore, we train AG-CNN with 70% of the dataset, but 80% are used in (Kumar et al., 2018; Rajpurkar et al., 2017). In nearly all the 14 classes, our method yields best performance. Only Rajpurkar *et al.* (Rajpurkar et al., 2017) report higher accuracy on *Hernia*. In all, the classification accuracy reported in this paper compares favorably against previous art.

Variant of training strategy analysis. Training three branches with different orders

Table 3.3 : Results of different training strategies.

Strategy	Global	Local	Fusion
GL_F	0.823	0.801	0.825
GLF	0.843	0.806	0.845
G_LF	0.841	0.809	0.843
G_L_F	0.841	0.817	0.868

influences the performance of AG-CNN. We perform 4 orders to train AG-CNN: 1) train global branch first, and then local and fusion branch together (G_LF); 2) train global and local branch together, and then fusion branch (GL_F); 3) train three branches together (GLF); 4) train global, local and fusion branch sequentially (G_L_F). Note that G_L_F is our three-stage training strategy. We train the AG-CNN with different training strategies. The experimental settings are same as Section 3.3.1. We present the classification performance of these training strategies in Table. 3.3.

AG-CNN yields better performance (0.868) with strategy of training three branches sequentially (G_L_F). When global branch is trained first, we perform the same model as the baseline in Table. 3.1. Training with G_L_F, AG-CNN obviously improves the baseline from 0.841 to 0.868. Compared with G_L_F, performance of AG-CNN (G_LF) is much lower because its the inaccuracy of local branch. When AG-CNN is trained with GL_F and GLF, it is inferior to G_L_F. Compared with training two or three branches (GL_F or GLF) together, training three branches in order (G_L_F) achieves much better performance. This is because that training global branch first could provide a relatively accurate discriminative region as the input of local branch. The performance of local branch is serious dependent on the global branch. From Table.3.3, we observe that a better performance in local branch leads to better performance in fusion branch. We infer

Table 3.4 : Results corresponding different statistics.

Statistic	Global	Local	Fusion
Max	0.8412	0.8171	0.8680
L1	0.8412	0.8210	0.8681

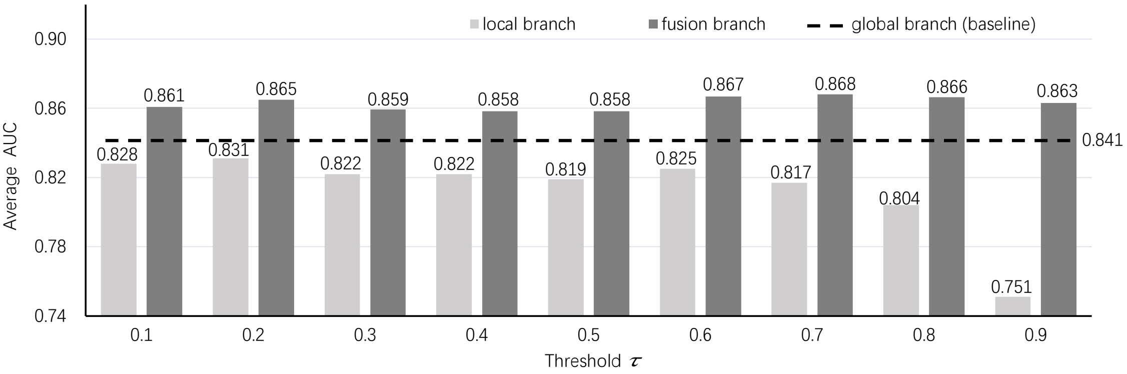


Figure 3.9 : Average AUCs for different settings of τ on the test set (ResNet-50 as backbone). Note that the results from global branch are our baseline.

that the performance of local branch is essential to enhance the whole framework.

Variant of heatmap analysis. In Table. 3.4, we report the performance of using different heatmap computing methods. Based on the same baseline, the performance is very close on both the local and fusion branch. It illustrates that different statistics result in subtle differences in local branch, but will not effect the classification performance significantly.

3.3.3 Parameter Analysis

We analyze the sensitivity of AG-CNN to the parameter consists in τ in Eq. 3.4, which defines the local region and affects the classification accuracy. Fig. 3.10 shows the average

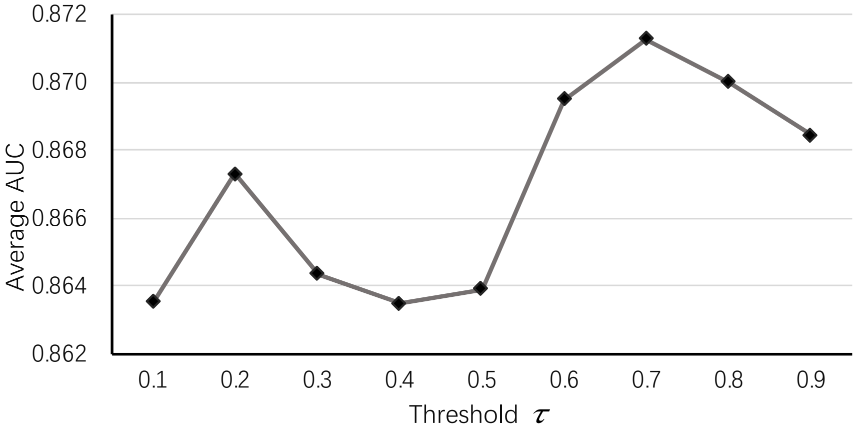


Figure 3.10 : Average AUC scores of AG-CNN with different settings of τ on the validation set (ResNet-50 as backbone).

AUC of AG-CNN over different τ on validation set. τ changes from 0.1 to 0.9. AG-CNN is not very sensitive to the threshold in the mask inference. The variance of the model performance is about 0.003 over the different τ. While τ is larger than 0.5, AG-CNN achieves much more stable and better performance (the average AUC is over 0.868), especially when τ is in [0.6, 0.8]. AG-CNN achieves the best performance when τ is setting as 0.7. Fig. 3.9 compares the average AUC of the global, local branch and fusion branch on the test dataset when ResNet-50 is used as backbone. When τ is small (*e.g.*, close to 0), the local region is close to the global image. In such cases, most of the entries in the attention heatmap are preserved, indicating that the cropped image patches are close to the original input. On the other hand, while τ is close to 1, *e.g.*, 0.9, the local branch is inferior to the global branch by a large margin (0.9%). Under this circumstance, most of the information in the global image is discarded but only the top 10% largest values in the attention heatmap are retained. The cropped image patches reflect very small regions. Unlike the local branch, AG-CNN is relative stable to changes of the threshold τ. When concentrating the global and local branches, AG-CNN outperforms both branches by at least 1.7% at $\tau = 0.4$ and 0.5. AG-CNN exhibits the highest AUC (>0.866) when τ

ranges between [0.6, 0.8].

3.4 Conclusion

This chapter proposes an attention guided convolutional neural network for chest X-ray image classification. Departing from previous works which merely rely on the global information, we propose to combining the global and the local cues to make diagnosis. An attention guided inference method is proposed to localize the most discriminative region in the global image. Extensive experiments demonstrate that combining both global and local cues yields state-of-the-art accuracy on the ChestX-ray14 dataset.

Chapter 4

Category-wise Residual Attenion Learning

4.1 Introduction

Commonly, CXR images are labeled with one or more pathologies, which makes the CXR image classification a multi-label problem. In the ChestX-ray14 dataset (Wang et al., 2017b), each image is annotated with multiple lung-related or heart-related pathologies. In the previous works, all the pathologies are equally treated in classifier learning. That is, when predicting the labels of each image, all pathologies are given the same weight. Furthermore, correlation essentially exists among the labels, *e.g.*, the presence of cardiomegaly additionally accompanies high risk of pulmonary edema. Most previous works focus on the multi-label setting on the disease space (or label space). Kumar *et al.*(Kumar et al., 2018) propose a boosted cascaded convolutional network framework which is similar to the classifier chains. Binary relevance and pairwise error loss function with the corresponding boosted cascaded structures are investigated in standard multi-label classification setting. Yao *et al.*(Yao et al., 2017) learn the dependencies of multiple diseases in the label space with a Long-short Term Memory Network (LSTM) (Hochreiter and Schmidhuber, 1997). The classification accuracy of (Kumar et al., 2018) or (Yao et al., 2017) is improved compared to the corresponding baseline method, which benefits from encoding the disease correlations in the label space. Therefore, exploring the dependency or correlation among labels could assist to strengthen the intrinsic relationship for some categories. However, considering an individual image, the uncorrelated labels may also introduce unnecessary noise and hinder the classifier from learning powerful features.

In this chapter, we present a category-wise residual attention learning (*CRAL*) framework for multi-label chest X-ray image classification. The proposed *CRAL* aims to mitigate the interference of uncorrelated classes and preserve correlations among the relevant classes at the same time. *CRAL* performs a category-wise residual attention mechanism to assign different weights to different feature spatial regions. It automatically predicts the attentive weights to enhance the relevant features and restrain the irrelevant features for a specific pathology. Figure 4.1 shows the architecture of *CRAL* framework. It consists of a feature embedding module and an attention learning module. The feature embedding module extracts high-level image features with a convolutional neural network. Attention module learns the normalized attention scores from the CNN features. By combining the channel-wise Hadamard product and element-wise sum operations, the high-level features and the attention scores are integrated into a residual attention block to classify the input image. The work of this chapter departs from the previous works in that we focus on reducing the obstruction of irrelevant features for one specific class while enhancing the relevant cues among all categories in the feature space. We show that CARL yields favorable performance compared with the state of the art.

Our contributions are summarized as follows:

- We propose a novel category-wise residual attention learning (*CRAL*) framework for multi-label chest X-ray image classification.
- *CRAL* benefits from both category-wise and residual attention learning. Residual attention learning advances in discriminative feature learning, and category-wise mechanism employs the correlations among pathologies to leverage the classification performance.
- We present the comprehensive experiment on the ChestX-ray14 dataset. Experimental results demonstrate that our framework yields superior performance over the state-of-the-art approaches.

4.2 The proposed method

In this section, we introduce the details of the proposed category-wise residual attention learning (*CRAL*) framework for the multi-label chest X-ray image classification. We will first describe the architecture of *CRAL* in Section 4.2.1. Then, the feature embedding module and the residual attention module are introduced in Section 4.2.2 and Section 4.2.3, respectively. We finally present the optimizing strategy in Section 4.2.4.

4.2.1 Architecture of CRAL

The architecture of *CRAL* is presented in Figure 4.1. It consists of a feature embedding module and an attention learning module. The feature embedding module learns the discriminative image features by a convolution neural network (CNN). The discriminative features are fed into the attention modules to learn the category-wise attention scores. And then they are used for adaptively assigning soft weights to different spatial positions of feature maps. Similar to (He et al., 2016; Wang et al., 2017a), we construct a residual attention architecture by adding the CNN feature and the corresponding weighted version. Finally, a binary classifier for each class is designed to classify the input image.

Multi-label Setup. We label each image with a C-dim vector $L = [l_1, l_2, ..., l_C]$ in which $l_c \in \{0, 1\}$. l_c represents whether the c^{th} pathology is presence or not, *i.e.*, 1 for presence and 0 for absence. C is the number of all pathologies in the dataset. If L is a zero vector, it means that none of all pathologies exists in the image.

4.2.2 Feature Embedding

Feature embedding module aims to extract a discriminative feature map $F \in R^{H \times W \times N}$ for each input image I by feeding it into a CNN model. Many deep learning-based methods have been proposed for this purpose. Here, we utilize ResNet-50 (He et al., 2016) or DenseNet-121 (Huang et al., 2017) network as the backbone. Next, we take ResNet-50 as an example to introduce the feature embedding module.

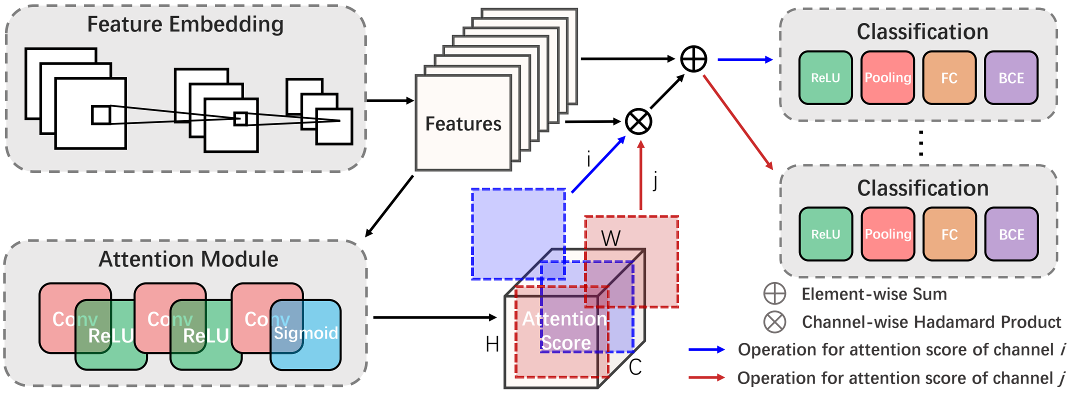


Figure 4.1 : Overview of the framework. There are two different attention mechanisms investigated in Section 4.2. Here, we take the first one *att1* as an example to illustrate the proposed framework. *CRAL* consists of two main modules. The feature embedding module is a CNN network which can be replaced by any network. In our experiment, we use ResNet-50 or Densenet-121 as the backbone. The normalized attention scores are obtained from the attention module. Attention scores contain *C* channels, and each channel corresponds to one category (highlighted with blue or red). By combining the channel-wise Hadamard product and element-wise sum operations, the high-level features and the attention scores are integrated into a residual attention block to classify the input image. Each class/disease is classified by a binary classifier in our model. "Pooling" represents a global average pooling layer. "FC" and "BCE" represent the fully connected layer and the binary cross entropy loss function, respectively.

The feature embedding module consists of five down-sampling residual blocks. Given a chest X-ray image I, the $H \times W$ feature map (for 224×224 input images) from layer "conv_5_relu" is used as input of attention module,

$$F = f_{cnn}(I; \theta_{cnn}), F \in \mathbb{R}^{H \times W \times N} \tag{4.1}$$

where θ_{cnn} is the parameters in feature embedding module, F is the feature maps from layer "conv_5_relu", N is the number of the feature channels. With DenseNet-121, we also extract the features from the "conv_5_relu" layer.

Another component of the proposed *CRAL* is the category-wise residual attention module which learns the discriminative spatial weight assignment scheme.

4.2.3 Category-wise Residual Attention Learning

Every image is semantically assigned one or more pathologies based on the multiple lesion regions. Although the positions of lesion areas are not provided, it is still expected that the model could pay attention to the relevant discriminative regions for classification. In this work, we focus on learning to predict such relevant regions for each class with attention mechanism under image-level supervisions. The attention maps are used to regularize the feature maps learned from feature embedding module. Basically, we expect to learn an attention score map whose values range from 0 to 1. The scores are leveraged to assign weights to different feature spatial regions for each pathology. The larger the attention score, the greater the weight is given to the corresponding position of the feature map, and thus the feature representation of the position is enhanced and vice versa. Therefore, the automatically predicted attention scores could aid to enhance the relevant features and restrain the irrelevant features for a specific pathology.

We investigate two different configurations of residual attention module which are denoted as $att1$ and $att2$, respectively. The architectures of residual attention are presented in Figure 4.2. *att1* consists of two 3×3 convolutional layers and each followed by a

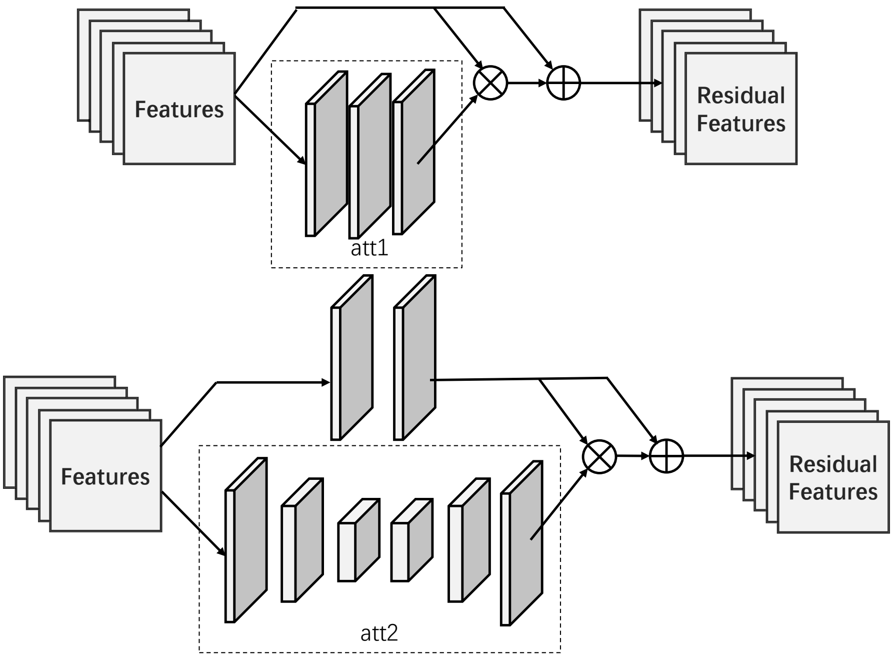


Figure 4.2 : Architecture of residual attention module (with *att1* and *att2*). *att1* consists of two 3 × 3 convolutional layers followed by ReLU, one 1 × 1 convolutional layer and one non-linear activation layer (Sigmoid). For *att2*, the input CNN features F are fed into the “hourglass” attention branch and a convolutional branch, respectively. Through the channel-wise Hadamard product and element-wise sum operations, a residual feature is formed by the learned features $\tilde{\boldsymbol{F}}$ and its weighted version $\boldsymbol{A} \odot \tilde{\boldsymbol{F}}$.

non-linear activation layer (ReLU), one 1×1 convolutional layers and a non-linear normalization layer (Sigmoid). The output is a C-channels attention scores corresponding to the C classes in the dataset. *att2* is similar to the hourglass structured attention proposed in (Wang et al., 2017a). It achieves to obtain a large receptive field by several max pooling layers among the residual blocks, and the global information is then expanded by a symmetrical up sample architecture. The last two convolutional layers are two consecutive 1×1 convolution layers. The last one outputs a C-channel attention score. Only one hourglass structured attention is stacked onto the last convolutional layer of feature embedding module. Except for the architectures, shown in Figure 4.2, we can see that another difference between two residual attention blocks is the feature maps are fed into another two residual blocks in *att2* while *att1* not. The residual attention with $att1$ also can be considered as identity mapping.

We formatively introduce the details of the attention module. For simplicity, we utilize att to represent either of the attention structures except for special situation. Given the CNN feature F, we aim to automatically predict label attention scores for each class,

$$Z = f_{att}(F; \theta_{att}), Z \in \mathbb{R}^{H \times W \times C} \tag{4.2}$$

where θ_{att} represents the parameters in attention module, Z is the unnormalized attention scores learned by f_{att} with each channel corresponding to one class. Z is then normalized with the sigmoid function to obtain the normalized attention scores A,

$$a_{i,j}^{c} = \frac{1}{1 + exp(-z_{i,j}^{c})}, A \in \mathbb{R}^{H \times W \times C} \tag{4.3}$$

where $a_{i,j}^{c}$ and $z_{i,j}^{c}$ represent the normalized and unnormalized attention scores at position (i, j) for c^{th} class, respectively. Intuitively, if the label c is tagged to the input image, the image regions related to it should be assigned with higher attention scores. Thus, the attention scores can be used to weight the CNN features for each class.

Afterwards, the CNN features are weighted by the attention scores. We take $att1$ as an example to illustrate the remain of the attention module in the following section.

The category-wise weighted CNN features are denoted as $V = \{V^1, V^2, ..., V^C\}$, where $V^c = \{v^{1,c}, v^{2,c}, ..., v^{N,c}\}$. Each channel $v^{n,c}$ of V^c is generated by channel-wise element-wise multiplication of each feature channel F^n with the attention score for one specific class a^c,

$$v_{i,j}^{n,c} = F_{i,j}^n \odot a_{i,j}^c, v^{n,c} \in \mathbb{R}^{H \times W} \tag{4.4}$$

where $\odot$ represents the Hadamard product. The weighted feature $v^{n,c}$ is more related to image regions corresponding to class c where n ranges from 1 to N.

However, naive attention module leads to obvious performance drop. This is because that the discriminative feature response values are weakened by the attention weights (range from 0 to 1). Therefore, similar to ideas in residual learning, we construct residual attention learning with the category-wise attention maps. Thus we combine the CNN features and the attended maps as

$$H_{i,j}^{n,c} = F_{i,j}^n + V_{i,j}^{n,c} = (\mathbf{1} + a_{i,j}^c) \odot F_{i,j}^n, \tag{4.5}$$

where $a_{i,j}^c$ ranges in [0,1], and it works as feature selectors which enhance discriminative features and suppress irrelevant features. **1** represents an all-ones matrix. Next, H^c is fed into a non-linear activation layer (ReLU) and a global average pooling layer (GAP). Specially,

$$\tilde{H^c} = max(0, H^c) \tag{4.6}$$

and

$$\bar{H}^{n,c} = \frac{1}{K} \sum_{i,j} \tilde{H}_{i,j}^{n,c}, \tag{4.7}$$

where K is the number of activation values in $\tilde{H}^{n,c}$. $\bar{H}^c = \{\bar{H}^{1,c}, \bar{H}^{2,c}, ..., \bar{H}^{N,c}\}$ is a N-dim vector. For $att2$, due to the CNN features from the feature embedding module are fed into two residual blocks, the F in Eq. 4.4 and Eq. 4.5 should be replaced by the new features $\tilde{F}$. We discuss these two attention mechanisms in Section 4.3.

Our residual attention module aims to learn the discriminative features for the multi-label chest X-ray image classification. The relationships between or within classes are implicitly presented in the high-level features which are automatically learned in the category-wise residual attention network.

4.2.4 Optimization

We define a binary classifier for each pathology in *CRAL* model. Note that the input of each classifier is not the same features. For the c^{th} class, the feature $\bar{H}^c$ is fed into a fully connected (FC) layer for classification,

$$\hat{H}^c = f_{cls}(\bar{H}^c; \theta^c), \tag{4.8}$$

where θ^c is the parameters of c^{th} classifier. Then a sigmoid layer is added to normalize the predicted confidence score $p(c|\hat{H}^c)$ of FC layer by

$$\tilde{p}(c|\hat{H}^c) = \frac{1}{1 + exp(-p(c|\hat{H}^c))}, \tag{4.9}$$

where $\tilde{p}(c|I)$ represents the probability score of I belonging to the c^{th} class, $c \in \{1, 2, ..., C\}$. The parameters in FC layers are denoted as $\theta_{fcs} = [\theta^1, \theta^2, ..., \theta^C]$. We optimize the parameters $W = [\theta_{cnn}, \theta_{att}, \theta_{fcs}]$ in *CRAL* by minimizing the binary cross-entropy (BCE) loss:

$$\mathcal{L}(W) = -\frac{1}{C}\sum_{c=1}^{C} l_c log(\tilde{p}(c|\hat{H}^c)) + (1 - l_c)log(1 - \tilde{p}(c|\hat{H}^c)), \tag{4.10}$$

where l_c is the ground truth of the c^{th} pathology. The *CRAL* can be trained end-to-end.

4.3 Experiment

This section evaluates the performance of the proposed *CRAL*. We first introduce the experimental dataset, evaluation metric, and the experimental settings. Section 4.3.3 discusses different attention mechanisms and demonstrates the performance of the proposed

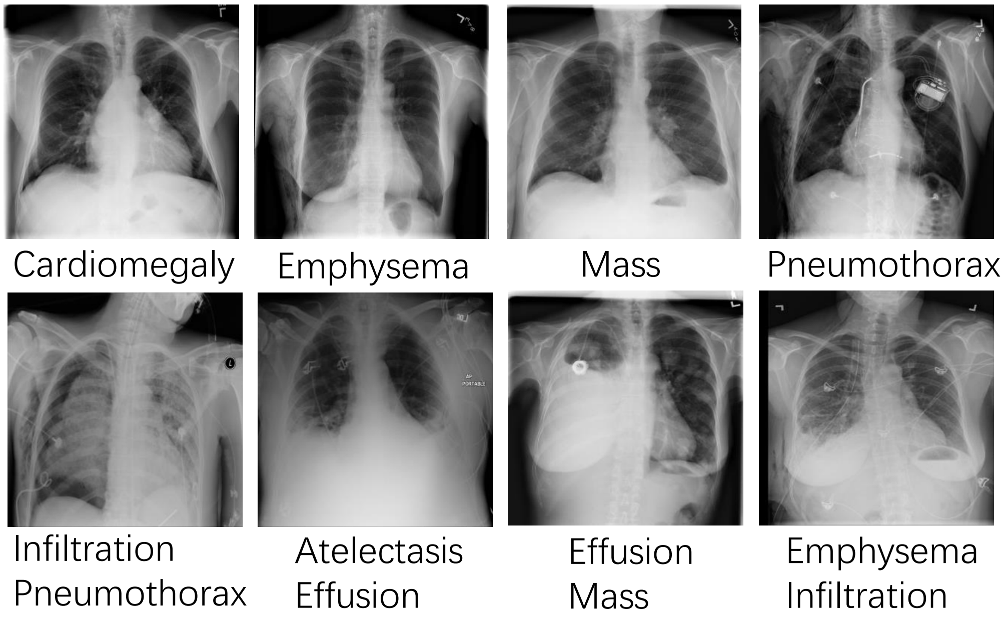


Figure 4.3 : Example images and the corresponding labels in the ChestX-ray14 dataset. Each image is labeled with one or more pathologies.

CRAL framework. The ablation study is presented to show the efficiency of *CRAL* in Section 4.3.4. At last, we visualize some feature heatmaps with *CRAL* and some classification results in Section 4.3.5.

4.3.1 Dataset and Evaluation Metric

Dataset. We evaluate the *CRAL* framework on a large scale chest X-ray dataset, ChestX-ray14, released by NIH (Wang et al., 2017b). It consists of 112,120 frontal-view X-ray images with 14 disease pathologies (Each image is assigned one or more pathologies. If there is no any pathology in an image, it is labeled as “No Finding”). Figure 4.3 shows some examples and the corresponding annotations in ChestX-ray14.

Evaluation Metric. In our experiment, we utilize the dataset split provided by (Wang et al., 2017b). There is no any patients overlap in train and test subsets. Each image is labeled with a one-shot vector $L = [l_1, l_2, ..., l_C]$, C is 14 in ChestX-ray14. Every element l_c represents the presence of the c^{th} pathology or not, *i.e.*, 1 for presence and 0 for absence. We use AUC score (the area under the ROC curve) of each pathology to measure the performance of *CRAL* framework.

4.3.2 Experimental Settings

For training, we perform data augmentation by resizing the input image to 256×256, randomly resized cropping to 224×224, and random horizontal flipping. The mean value of ImageNet is subtracted from the image. We optimize the network by SGD with a mini-batch size of 64 and train 30 epochs. The learning rate starts from 0.01 and is divided by 10 after 20 epochs. We use a weight decay of 0.0001 and a momentum of 0.9. During testing, the image is also resized to 256×256, and then center cropping is performed to obtain an image of size 224×224. The ImageNet mean value is also subtracted. The *CRAL* framework is implemented with Pytorch (Paszke et al., 2017).

4.3.3 Evaluation

We evaluate our method on the ChestX-ray14 dataset. ResNet-50 (He et al., 2016) and DenseNet-121 (Huang et al., 2017) are used as the basic backbone in feature embedding module. The corresponding AUC and ROC curves are presented. We first showcase the performance of different attention module structures under the *CRAL* framework, and then compare *CRAL* with the state-of-the-art methods.

Comparison with different attention structures. Firstly, we evaluate the attention mechanism $att1$ and $att2$ to validate the effectiveness of the proposed *CRAL* framework. The results are summarized in Table 4.1, Table 4.2 and Figure 4.4.

Both *att1* and *att2* improve the classification performance on the chestX-ray14 dataset. With ResNet-50 as the backbone, *att1* is little superior than *att2* (the average AUC scores over 14 pathologies are 0.814 and 0.810). With *att1*, AUC scores of “Fibrosis” are higher than *att2* by 0.013. For “Consolidation” and “Fibrous”, ResNet-50 with *att1* achieves

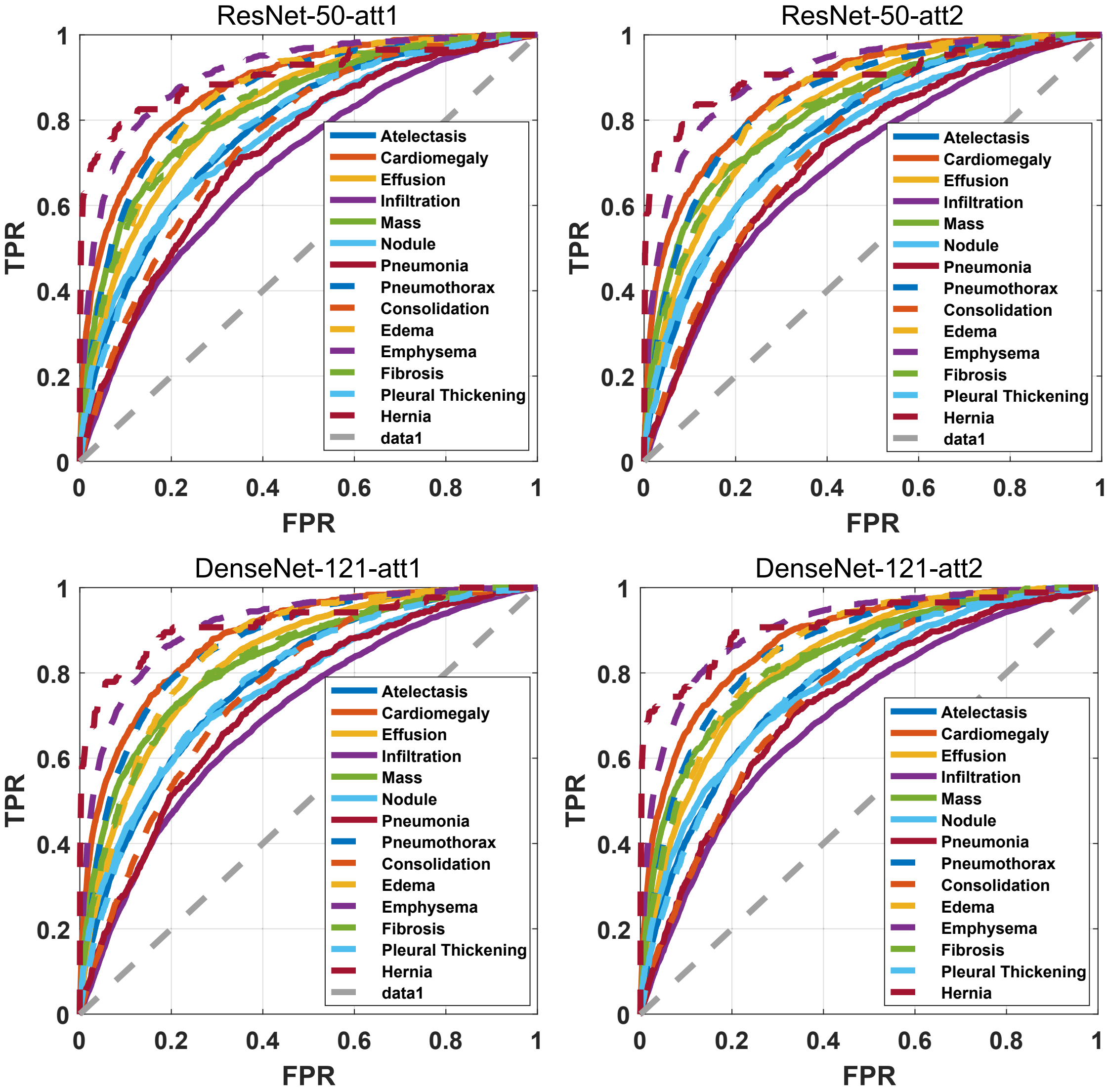


Figure 4.4 : ROC curves of four combinations of CNN backbones and attention mechanisms (ResNet-50-att1, ResNet-50-att2, DenseNet-121-att1, and DenseNet-121-att2) over the 14 pathologies. The corresponding AUC scores are given in Table. 4.1.

the highest AUC scores (0.758 and 0.832). The AUC scores of “Mass”, “Pneumonia” and “Pneumothorax” with *att1* exceeds about 0.005 compared with *att2*. While with DenseNet-121 as backbone, *CRAL* shows similar performance on 14 pathologies shown in Table 4.1 and Figure 4.4. The AUC scores of 9 pathologies with *att1* and another one pathology with *att2* achieve the state of the art.

Comparison with the state-of-the-art methods. Some previous methods, like (Yao et al., 2017), (Kumar et al., 2018) or (Rajpurkar et al., 2017), train and test using different dataset split strategies. For a fair comparison, we only compare the methods which utilize this public available split list provided by (Wang et al., 2017b) with image-level supervision. We evaluate *CRAL* and compare it with state-of-the-art methods on the ChestX-ray14 dataset. The results are summarized in Table. 4.1 and Figure. 4.4. Wang *et al.*(Wang et al., 2017b) integrate the classification and localization tasks into a unified framework. The localization is complemented based on the features by the image-level supervised learning. Guendel *et al.*(Guendel et al., 2018) propose a location aware Dense Network (DNetLoc), which incorporates both high-resolution image data and spatial information for pathology classification. DenseNet is used as the backbone network in DNetLoc which is same as ours. The main differences between (Guendel et al., 2018) and our method are in two folds: 1) DNetLoc achieves the high-resolution by inserting two convolutional layers with stride before the DenseNet, while ours focuses on improving the feature representation by the proposed category-wise residual attention mechanism. And 2) DNetLoc introduces the extra large-scale dataset with the disease position information to further improve the recognition performance while ours does not utilize any other extra data. Yao *et al.*(Yao et al., 2018) achieve the chest X-ray image classification and localization with a multiple resolutions setting. Li *et al.*(Li et al., 2018) utilize additional lesion area annotation as supervision. Tang *et al.*(Tang et al., 2018) progressively learn an attention-guided curriculum to identify the pathologies and the attributes mining from radiology reports are used. Shen *et al.*(Shen and Gao, 2018) combine the routing-by agreement mechanism and the deep convolutional neural network to achieve such goals.

Compared with these methods, this paper contributes a new state of the art: average AUC is 0.816. *CRAL* largely exceeds the previous works a large gap, especially (Wang et al., 2017b) and (Li et al., 2018) with 7.1% and 7.7%. With DenseNet-121, it surpasses the previous state of the art (Guendel et al., 2018) nearly 1%. *CRAL* achieves the state of the art on half of 14 pathologies. The scores of the other three pathologies "Nodule", "Pneumonia" and "Infiltration" are also competitive compared with the current highest scores (0.773 *vs.* 0.777, 0.729 *vs.* 0.731 and 0.702 *vs.* 0.709). More importantly, the AUC scores of some pathologies, *e.g.*, *Pneumothorax*, *Pleural Thickening*, *Edema*, or *Hernia*, are higher than (Guendel et al., 2018) about 2% (DenseNet-121 with *att2*). The ROC curves of 14 pathologies with ResNet-50 and DenseNet-121 are presented in Figure 4.4. In all, the classification performance reported in this paper compares favorably against previous methods.

Table 4.1 : Comparison results of various methods on ChestX-ray14. We compute the AUC score of each class and the average AUC scores across the 14 diseases. For each column, the best results are highlighted in bold.

Method	CNN	Atel	Card	Effu	Infi	Mass	Nodu	Pne1	Pne2	Cons	Edem	Emph	Fibr	PT	Hern	Mean
(Wang et al., 2017b)	R-50	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
(Guendel et al., 2018)	D-121	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
(Yao et al., 2018)	*	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
(Li et al., 2018)	R-50	0.727	0.836	0.789	0.672	0.776	0.696	0.649	0.808	0.720	0.806	0.888	0.771	0.737	0.693	0.755
(Li et al., 2018)	D-121	0.728	0.848	0.782	0.645	0.747	0.702	0.632	0.802	0.727	0.823	0.757	0.763	0.735	0.653	0.739
(Shen and Gao, 2018)	–	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
(Tang et al., 2018)	–	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
CRAL (att1)	R-50	0.779	0.879	0.824	0.694	0.831	0.766	0.726	0.858	0.758	0.850	0.909	**0.832**	0.778	0.906	0.814
CRAL (att2)	R-50	0.777	0.875	0.826	0.695	0.825	0.765	0.720	0.852	0.751	0.848	0.905	0.819	0.777	0.908	0.810
CRAL (att1)	D-121	**0.781**	0.883	**0.831**	0.697	0.830	0.764	0.725	**0.866**	0.758	**0.853**	**0.911**	0.826	**0.780**	**0.918**	**0.816**
CRAL (att2)	D-121	**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**

* The 14 pathologies are Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, Pneumothorax, Consolidation, Edema, Emphysema, Fibrosis, Pleural Thickening and Hernia, respectively. * represents that the combination of ResNet and DenseNet is used in Yao *et al.* (Yao et al., 2018). – represents the network used in the corresponding reference is not illustrated.

4.3.4 Ablation Study

To evaluate the effectiveness of the components of residual attention module, we conduct additional ablation experiments on the ChestX-ray14 dataset. *CRAL* is performed with ResNet-50 and DenseNet-121 as the backbone, combining *att1* and *att2*, respectively. We remove each component in *CRAL* at a time, including the category-wise operation, residual operation, and the whole attention module. Without attention (*w/o attention*) is considered as our baseline. Under the condition of without category setting, all the categories are weighted by the same attention scores. And "w/o residual" represents that only the features weighted by the category-wise attention scores are used to learn classifier. The average AUC scores are presented in Table 4.2. The performance of *CRAL* over 14 pathologies is reported in Table 4.1.

CRAL constantly improves the baseline nearly 1% (0.8136 *vs.* 0.8034, 0.8157 *vs.* 0.8056) with either ResNet-50 or DenseNet-121 as backbone. First, removing the whole residual attention, the remaining models with ResNet-50 and DenseNet-121 have AUC scores of 0.8034 and 0.8056, respectively. It is inferior to the full model. The performance drop is approximately 1%. It illustrates that the residual attention module is important for enhancing the relevant features but reducing the obstructions of irrelevant features. Second, after removing the "residual" configuration in *CRAL*, the performance drops significantly from 0.067 to 0.01, but it still superior to its corresponding baseline. Besides, the removal of "category" makes the performance drop slightly compared with the "residual" setting. We summarize that the proposed *CRAL* improves the performance of multi-label chest X-ray image classification.

4.3.5 Qualitative results

We visualize some feature heatmaps and classification results shown in Figure 4.5 and Figure 4.6, respectively. The heatmap is generated by two steps: we first take the absolute value of the feature values at each position from a specific layer (the *conv_5* layer of

Table 4.2 : Comparison of ablation study with different experimental setting. The average AUC scores are reported. *CRAL* is considered as the "full" model. We remove the category-wise, residual operation and the whole attention component at a time. They are denoted as *w/o category*, *w/o residual* and *w/o attention*, respectively. Two different attention mechanisms in *CRAL* are performed in the experiment.

	Backbone: R-50		Backbone: D-121	
Method	Att1	Att2	Att1	Att2
CRAL	**0.8136**	**0.8102**	**0.8157**	**0.8157**
w/o category	0.8114	0.8093	0.8135	0.8136
w/o residual	0.8069	0.8052	0.8069	0.8073
w/o attention	0.8034	0.8034	0.8056	0.8056

ResNet-50), and then count the maximum values along feature channels. In Figure 4.5, we observe that the discriminative regions of the images are activated. It demonstrates that the *CRAL* could learn to focus on the lesion areas which leads to accurately recognize the pathologies. In Figure 4.6, the top-8 probability scores are presented for each sample. The ground truth labels are highlighted in red or blue. We see that large gaps generated by the scores of true pathologies and other pathologies, *e.g.*, the predicted score of "Cardiomegaly" (row 1, column 3) is 0.8873 which is about 40 times of "Nodule" (0.0265). Only for several special cases (highlighted in blue), *CRAL* does not accurately recognize the pathologies.

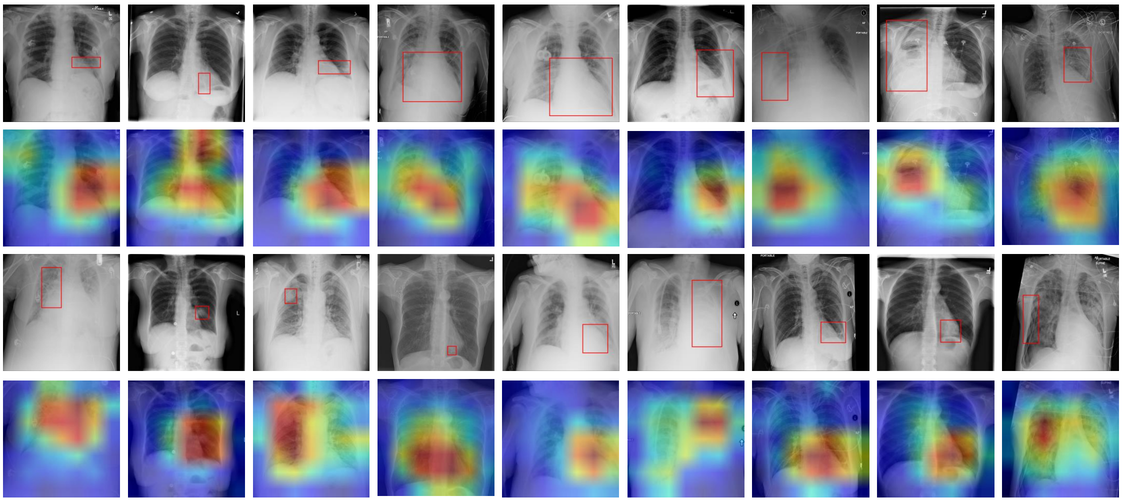


Figure 4.5 : Examples of heatmaps generated from the learned features (from ResNet-50). The ground truth bounding boxes provided by (Wang et al., 2017b) are annotated on the original images. Note that the heatmaps are zoomed to the same size as the input images, and the heatmaps may be a few difference due to the usage of random cropping in testing.

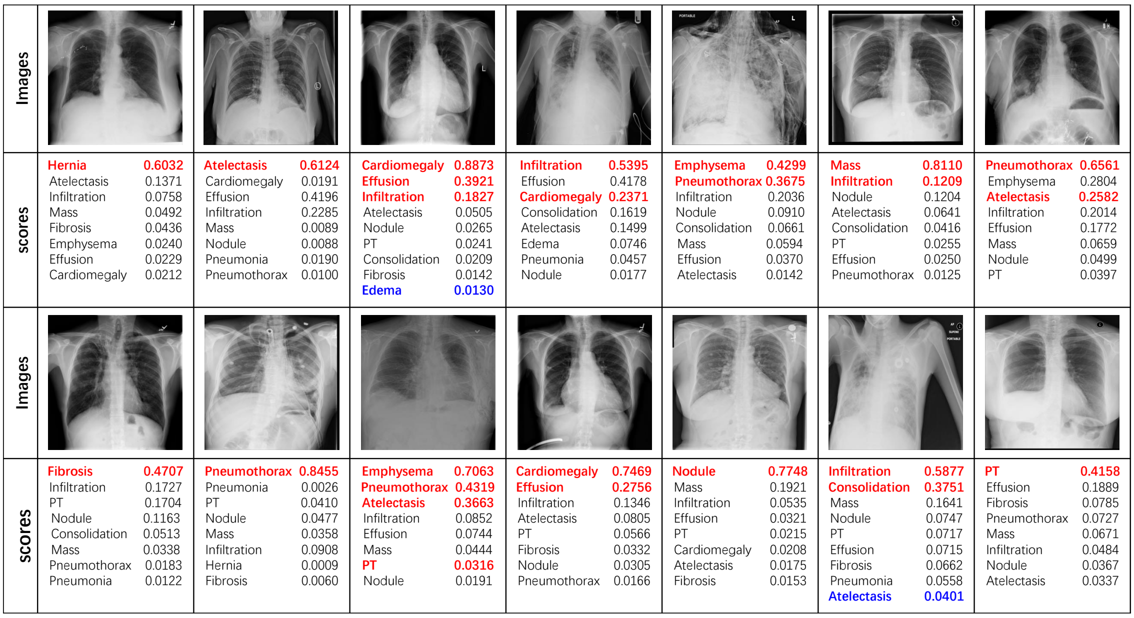


Figure 4.6 : Examples of classification results. We present the top-8 predicted categories and the corresponding probability scores. The ground truth labels are highlighted in red or blue.

4.4 Conclusion

This chapter proposes a category-wise residual attention learning framework for the multi-label chest X-ray image classification. The proposed framework learns the discriminative features for multi-label classification end-to-end. Depart from the previous works, we perform the category-wise attention to induce the obstruction from irrelevant classes and enhance the weights within the relevant classes. Extensive experiments illustrate that the category-wise residual attention mechanism is efficient to classify the chest X-ray images. Experiments on the ChestX-ray14 dataset demonstrate the effectiveness of the proposed method.

Chapter 5

Discriminative Feature Learning

This chapter focuses on learning discrimative features for multiple disease classification in chest X-ray images. Different from the generic image classification task, a robust and stable CXR image analysis system should consider the unique characteristics of CXR images. Particularly, it should be able to: 1) automatically focus on the disease-critical regions, which usually are of small sizes; 2) adaptively capture the intrinsic relationships among different disease features and utilize them to boost the multi-label disease recognition rates jointly. We introduce a two-branch architecture, named *ConsultNet*, to achieve those two purposes simultaneously. *ConsultNet* consists of two components. First, an information bottleneck constrained feature selector extracts critical disease-specific features according to the feature importance. Second, a spatial-and-channel encoding based feature integrator enhances the latent semantic dependencies in the feature space. *ConsultNet* fuses these discriminative features to improve the performance of thorax disease classification in CXRs. Experiments conducted on the ChestX-ray14 and CheXpert dataset demonstrate the effectiveness of the proposed method.

5.1 Introduction

Developing a stable and robust computer-aided disease analysis system is critical to assist disease diagnosis and treatment. For the thorax disease classification problem, such a system is necessary to focus on the potential lesion areas and suppress the noise introduced by the irrelevant regions. Additionally, exploring the intrinsic correlations of multiple diseases is also beneficial to improve the performance of the computer-aided system.

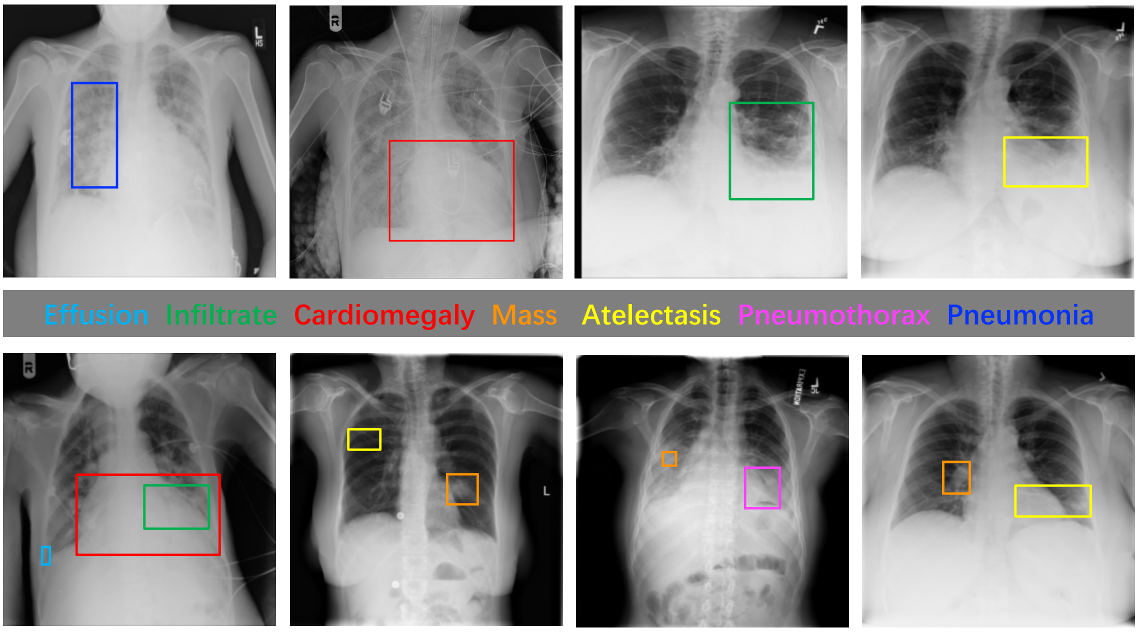


Figure 5.1 : Examples of lesion areas on the ChestX-ray14 dataset. The first row presents some chest X-ray images with lesion areas, which are small compared to the global ones. The second row shows multiple pathologies existing in an image, which means the corresponding patient suffers from various diseases in a period. The disease existing in each bounding box corresponds to the pathology name with same color in the middle row.

Chest X-ray image classification suffers from a large amount of disease-irrelevant regions. As shown in Fig. 5.1, in most cases, most regions in given images contain healthy tissues. Therefore, they provide little useful knowledge for diagnosis and lead to unnecessary computation cost. Because of the very relative small lesion areas of some pathologies in Fig. 5.1 (such as “Effusion” or “Mass”) or some special cases showing in Fig. 5.5 (the second row), it is necessary to exclude the interference of disease-uncorrelated regions. Most of the existing works try to solve this problem from the aspect of introducing additional local-aware or medical report information (Tang et al., 2018; Li et al., 2018; Guendel et al., 2018; Guan et al., 2020). Guan *et al.*(Guan et al., 2020) and Guendel *et al.*(Guendel et al., 2018) propose to localize the region of interests and combine it with the global image to classify the chest X-ray image. Li *et al.*(Li et al., 2018) jointly implement the disease identification and localization under the supervision of image-level labels and limited lesion area location information. Combining the image-level labels and severity-level attributes mined from radiology reports, Tang *et al.*(Tang et al., 2018) propose to identify the pathology and localize the lesion areas at the same time. The additional information could help to focus on the region of interests, but it also introduces more extra computational or data annotation cost. It is better to introduce a scheme to make the computer-aided system automatically focus on the regions of interest but not add any other burden.

Besides, it is not rare to find that one patient is suffering from more than one disease at the same time in real scenarios. Fig. 5.1 (the second row) shows several examples of multiple diseases presented in one image. The lesion areas may occur in either nearby or separate regions. Because of the essentially existing correlations among the diseases, it is necessary to capture the correlations of multiple diseases to classify the CXR image. Kumar *et al.*(Kumar et al., 2018) and Yao *et al.*(Yao et al., 2017) propose to boost the ConvNets from the aspect of multi-label dependencies on the network output space, which is limited with the static label setting. Most of the previous works focus on the relations between different pathologies in the label space. In this paper, we explore the latent semantic dependencies of multiple diseases in the feature space. Thus, more discriminative features are expected to be learned for multi-label CXR image classification task.

Last but not least, due to the high appearance similarity of chest X-ray images, the diagnosis is difficult because the tremendous inter-class similarity may affect the discriminative feature learning and reduce the classification performance. This combination of above problems makes the task of chest X-ray image classification challenging even for the powerful deep learning algorithms.

In this paper, we develop a novel framework, named ***ConsultNet***, to solve the aforementioned issues. The framework is presented in Fig. 5.2. *ConsultNet* consists of two modules. First, we introduce a spatial-wise and channel-wise based attention mechanism into the Variational Information Bottleneck (VIB) to enforce the network to select critical, disease-specific features for chest X-ray image classification. We name this principle as Variational Selective Information Bottleneck (VSIB). VSIB derives from the observation that a uniform information constraint may not promote the network to focus on as many disease-specific features as possible. We achieve the above purpose according to the feature importance under an information bottleneck constraint of latent feature representation. It constrains that as much as disease-specific information passing the bottleneck is reserved. Thus, the features of disease-irrelevant regions are excluded. VSIB does not need extra bounding box annotations. It would not introduce massive computational costs, either. We name this module as *Feature Selector*. Second, a Spatial-and-Channel Encoding (SCE) module is proposed to model the latent semantic dependencies of multiple diseases in the feature space. The SCE module serves as a *Feature Integrator* to strengthen long-range relationships of features in both spatial and channel dimensions. By this development, *ConsultNet* can not only distill the most critical information selectively from CXR images but also model the feature semantic correlations explicitly under a unified framework.

ConsultNet learns discriminative features for CXR image classification from different views. The VSIB module filters the critical features for classification, and the SCE module encodes the feature semantic dependencies in the feature space. As shown in Fig. 5.6, we experimentally observe that these two modules function as two collaborative feature learner to make an accurate diagnosis. Fusing the multi-view features could provide more information and benefit to disease recognition. Moreover, we address the more substantial inter-class appearance similarity of chest X-ray images by regularizing the *ConsultNet* with a pairwise confusion strategy (Dubey et al., 2018), which can enforce the *ConsultNet* to forget the patient-specific features but remember the disease-specific features.

We summarize the contributions of this work as follows:

- We propose a two-branch *ConsultNet* that collaboratively learn discriminative features for CXR image classification.
- A novel variational selective information bottleneck (VSIB) principle is proposed to induce the *ConsultNet* to pay more attention to the disease-correlated regions and preserve more discriminative features.
- We propose to strengthen the semantic dependencies of multi-disease features in the feature space with a Spatial-and-Channel Encoding module.
- To address the inter-class sample similarity problem in chest X-ray images, we propose to train *ConsultNet* with a pairwise confusion strategy.

5.2 Methodology

5.2.1 Problem Settings and Motivation

We focus on the problem of discriminative feature learning in chest X-ray image classification. Given a dataset $(\boldsymbol{X}, \boldsymbol{Y}) = \{\boldsymbol{x}_i, \boldsymbol{y}_i\}, i \in \{1, 2, ..., \mathcal{N}\}$, where $\boldsymbol{x}_i$ is a chest X-ray image, $\boldsymbol{y}_i$ is a vector denoted as $[y_i^1, y_i^2, ..., y_i^{\mathcal{C}}]$, in which each element y_i^c represents the presence of the c^{th} pathology or not, *i.e.*, 1 for presence and 0 for absence. $\mathcal{C}$ is the number of pathologies. $\mathcal{N}$ is the number of samples/images.

First, our target is to learn a model $\mathcal{M}$ (with parameter $\boldsymbol{W}$) that can encode a compact, disease-specific feature representation $\boldsymbol{z}_i$ from the given sample $\boldsymbol{x}_i$. $\boldsymbol{z}_i$ is informative over the class $\boldsymbol{y}_i$. For the convenience of further discussions, we divide $\mathcal{M}$ into a feature encoder E and a decoder D and thus $\boldsymbol{z}_i = E(\boldsymbol{x}_i)$. The output of D is a categorical distribution $D(\boldsymbol{z}_i) = p(\hat{\boldsymbol{y}}_i|\boldsymbol{z}_i)$. Then we could minimize

$$\mathcal{L}(\boldsymbol{W}) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim p(\boldsymbol{X},\boldsymbol{Y})}[\ell(\boldsymbol{x}, \boldsymbol{y})], \tag{5.1}$$

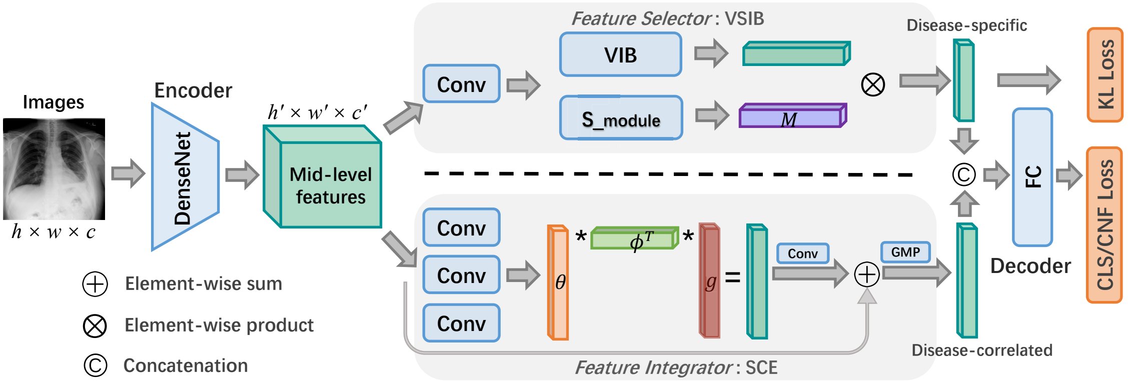


Figure 5.2 : Overview of the proposed *ConsultNet*. The *ConsultNet* consists of an Encoder, a Feature Selector, a Feature Integrator, and a Decoder. Given an image, we first feed it into the Encoder and obtain a mid-level feature representation. Then we learn the disease-specific and disease-correlated features by a VSIB based Feature Selector and an SCE based Feature Integrator, respectively. At last, both of them are concentrated together to classify the input image. Note that the "Conv", "VIB", "VSIB","SCE", "GMP" and "FC" represent the convolutional layer, variational information bottleneck, spatial-channel encoding, global max pooling layer and fully connected layer respectively.

where $\mathbb{E}[\cdot]$ represents statistical expectation and $\ell(\cdot)$ is binary cross entropy:

$$\ell(\boldsymbol{W}) = -\frac{1}{\mathcal{C}} \sum\nolimits_{c=1}^{\mathcal{C}} y_i^c log(p(\hat{y}_i^c|\boldsymbol{x})) + (1 - y_i^c) log(1 - p(\hat{y}_i^c|\boldsymbol{x})). \tag{5.2}$$

Chest X-ray image classification suffers from a large number of noisy regions outside the lesion area and the lack of an explicit mechanism to capture the relationships among multiple diseases. Accordingly, we propose the *ConsultNet* to achieve these purposes at the same time. The architecture of *ConsultNet* is shown in Fig. 5.2. Given an image, we extract its mid-level features with a DenseNet (Huang et al., 2017). Next, on the one hand, we extract the disease-specific features with a variational selective information bottleneck as a constraint. On the other hand, we characterize the dependencies among multiple diseases to yield disease-correlated features. We denote these two modules as *Feature Selector* and *Feature Integrator*, respectively. At last, both of the above features are concatenated together to classify the CXR image.

We will introduce the variational selective information bottleneck based *Feature Selector* in Sec. 5.2.2 and the spatial-and-channel *Feature Integrator* in Sec. 5.2.3, respectively. The details of optimizing strategy are presented in Sec. 5.2.4.

5.2.2 Variational Selective Information Bottleneck

From the view of supervised learning, the goal of a variational information bottleneck (VIB) is to learn a representation $\boldsymbol{Z}$, which is predictive of the label $\boldsymbol{Y}$ while encoding only a small amount of information from the input $\boldsymbol{X}$ (Alemi et al., 2017; Tishby and Zaslavsky, 2015). The latter is equal to bound the mutual information between $\boldsymbol{X}$ and $\boldsymbol{Z}$ to a specific threshold: $I(\boldsymbol{X}, \boldsymbol{Z}) \leq I_c$, where I_c is the information constraint. This suggests the following objective:

$$\max\nolimits_{\boldsymbol{W}} I(\boldsymbol{Z}, \boldsymbol{Y}; \boldsymbol{W}) \quad s.t. \quad I(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{W}) \leq I_c. \tag{5.3}$$

We first review the vanilla variational information bottleneck (VIB) principle in deep learning. Built on the recently developed information theoretic objectives for deep neural networks (Alemi et al., 2017; Tishby and Zaslavsky, 2015), such information constraint is performed on the encoder E as a *Feature Selector* and results in the following objective:

$$\begin{aligned} \mathcal{L}_{cls}(\boldsymbol{W}) &= E_{(\boldsymbol{x},\boldsymbol{y})\sim p(\boldsymbol{X},\boldsymbol{Y})}[\ell(\boldsymbol{x},\boldsymbol{y})] \\ s.t. \quad & E_{\boldsymbol{x}\in p(\boldsymbol{x})}(KL[E(\boldsymbol{z}|\boldsymbol{x})||r(\boldsymbol{z})]) \leq I_c, \end{aligned} \tag{5.4}$$

which $r(\boldsymbol{z})$ is a approximated prior margin distribution of $\boldsymbol{z}$, which is modeled with a standard Gaussian. Suppose we utilize the encoder E of the form $E(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}|f_e^{\boldsymbol{\mu}}(\boldsymbol{x}), f_e^{\boldsymbol{\Sigma}}(\boldsymbol{x}))$, where f_e is a DenseNet (Huang et al., 2017) backbone which outputs both the mean $\boldsymbol{\mu}$ of $\boldsymbol{z}$ as well as the covariance matrix $\boldsymbol{\Sigma}$. The reparameterization trick (Kingma and Welling, 2013) can be written with a $\boldsymbol{z} = \boldsymbol{\mu} + \boldsymbol{\Sigma}\epsilon$, where ϵ is auxiliary noise variable $\epsilon \sim \mathcal{N}(0, 1)$. Kullback-Leibler (KL) divergence measures an analytic of $E(\boldsymbol{z}|\boldsymbol{x})$ and $r(\boldsymbol{z})$. The large value of Kullback-Leibler (KL) divergence measures an analytic of $E(\boldsymbol{z}|\boldsymbol{x})$ and $r(\boldsymbol{z})$. Therefore, enforcing the KL divergence lower than a threshold I_c could lead to learn a compact, disease-specific feature representation. Eq. 5.4 equals to minimize the objective function

$$\begin{aligned} \mathcal{L}_{cls}(\boldsymbol{W}) &= E_{(\boldsymbol{x},\boldsymbol{y})\sim p(\boldsymbol{X},\boldsymbol{Y})}[\ell(\boldsymbol{x},\boldsymbol{y})] \\ &+\beta(E_{\boldsymbol{x}\in p(\boldsymbol{x})}(KL[E(\boldsymbol{z}|\boldsymbol{x})||r(\boldsymbol{z})]) - I_c), \end{aligned} \tag{5.5}$$

where β is a Lagrange multiplier, $\beta \geq 0$.

Apart from the vanilla VIB, we propose a variational selective information bottleneck (VSIB) which derives from the observation that a uniform information constraint may not promote the network to focus on as many disease-specific features as possible. As shown in Fig. 5.4 (the first two rows), some separate disease-unrelated regions are localized by VIB. The feature space will be disturbed by such bad cases. Moreover, as an example of "Pneumonia" in the third row of Fig. 5.4, VIB misses most of the lesion areas. However, we expect that the network could focus on the disease-specific regions and learn as many discriminative features as possible to classify the chest X-ray image.

To achieve the above purposes, we propose to introduce a selective mechanism to the VIB constraint. As shown in Fig. 5.2, the selective module (*S_module*) is utilized to

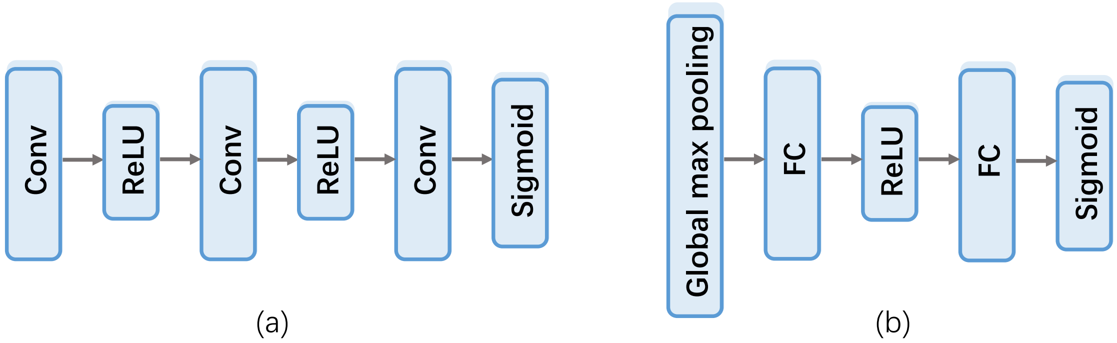


Figure 5.3 : The architectures of *S_module*. (a) and (b) represent the S_s and S_c submodules, respectively.

predict an importance matrix $\boldsymbol{M}$ to specify the importance of each element of the input feature. Specifically, we adopt two submodules (*Ss* and *Sc*) to character $\boldsymbol{M}$. The architectures of these two submodules are presented in Fig. 5.3. *Ss* (with parameters $\boldsymbol{W}_{ss}$) consists of two 3×3 convolutional layers, one 1×1 convolutional layer and a Sigmoid layer, acting on the spatial dense map and aiming to detect the spatial importance of features. *Sc* (with parameters $\boldsymbol{W}_{sc}$) is utilized to capture the importance of feature between channels, which comprises two fully connected layers and a Sigmoid layer. Before *S_module*, we add a 1×1 convolutional layer (with parameters $\boldsymbol{W}_t \in \mathbb{R}^{H \times W \times C}$, where H and W represent the spatial dimensions and C represents the channel dimension.) to transform the previous features as the input of VSIB. $\boldsymbol{M}$ is defined as

$$\boldsymbol{M} = \boldsymbol{Z}\boldsymbol{W}_t \otimes (\boldsymbol{W}_{ss} \odot \boldsymbol{W}_{sc}) \tag{5.6}$$

where $\otimes$ denotes the element-wise product. $\odot$ represents the matrix expansion and element-wise multiplication. It first expands the dimension of the spacial importance $\boldsymbol{W}_{ss} \in \mathbb{R}^{H \times W}$ and the channel-wise importance $\boldsymbol{W}_{sc} \in \mathbb{R}^{C}$ to $\mathbb{R}^{H \times W \times C}$. And then merges $\boldsymbol{W}_{ss}$ and $\boldsymbol{W}_{sc}$ into the elements-wise importance of the feature tensor by element-wise multiplication. To achieve a selective information constraint according to the feature importance, we perform the VSIB by weighting the VIB constraint adaptively with $(\mathbf{1}-\boldsymbol{M})$:

$$\begin{aligned}\mathcal{L}_{cls}(\boldsymbol{W}) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim p(\boldsymbol{X},\boldsymbol{Y})}[\ell(\boldsymbol{x},\boldsymbol{y})]+ \\ \beta\mathbb{E}_{\boldsymbol{x}\in p(\boldsymbol{X})}(\mathbf{1}-\boldsymbol{M})\otimes KL[E(\boldsymbol{z}|\boldsymbol{x})||r(\boldsymbol{z})].\end{aligned} \tag{5.7}$$

Finally, the features generated from *Feature Selector*, denoted as $\tilde{\boldsymbol{Z}}_{vsib}$, can be modeled as

$$\tilde{\boldsymbol{Z}}_{vsib} = \boldsymbol{M}\otimes\boldsymbol{Z}\boldsymbol{W}_t + \boldsymbol{Z}\boldsymbol{W}_t. \tag{5.8}$$

In Fig. 5.4(c), we present the feature heatmaps generated by VSIB. With VSIB as the feature constraint, the activated regions in the heatmaps are closer to the true lesion areas compared with VIB.

Generally, the parameters of this module are denoted as $\boldsymbol{W}_{vsib} = [\boldsymbol{W}_t, \boldsymbol{W}_{ss}, \boldsymbol{W}_{sc}]$.

5.2.3 Spatial-and-Channel Encoding

Recent approaches about multi-label chest X-ray image classification task focus on exploiting the disease dependencies in the label space (Kumar et al., 2018; Yao et al., 2018). However, the latent semantic dependencies are still not explored. In this work, inspired by (Yue et al., 2018), we introduce a *Feature Integrator* to encode the latent semantic dependencies of multiple diseases in the feature space. It computes the long-range response at a position as a weighted sum of the features between any positions of any channels. We name this module as Spatial-and-Channel Encoding (SCE). The SCE module is utilized to capture the intrinsic relationships among the features of multiple diseases and improve the disease recognition performance. SCE encodes the latent semantic dependencies of multiple diseases in the feature space. The semantic correlations between the different disease features are enhanced, and disease-uncorrelated features are expected to be inhibited in this operation at the same time.

Given a feature tensor $\boldsymbol{Z}\in\mathbb{R}^{N\times C}$ generated from the encoder network E (with parameter $\boldsymbol{W}_E$), *C* is the number of channels and *N* is the sum of spatial positions over a

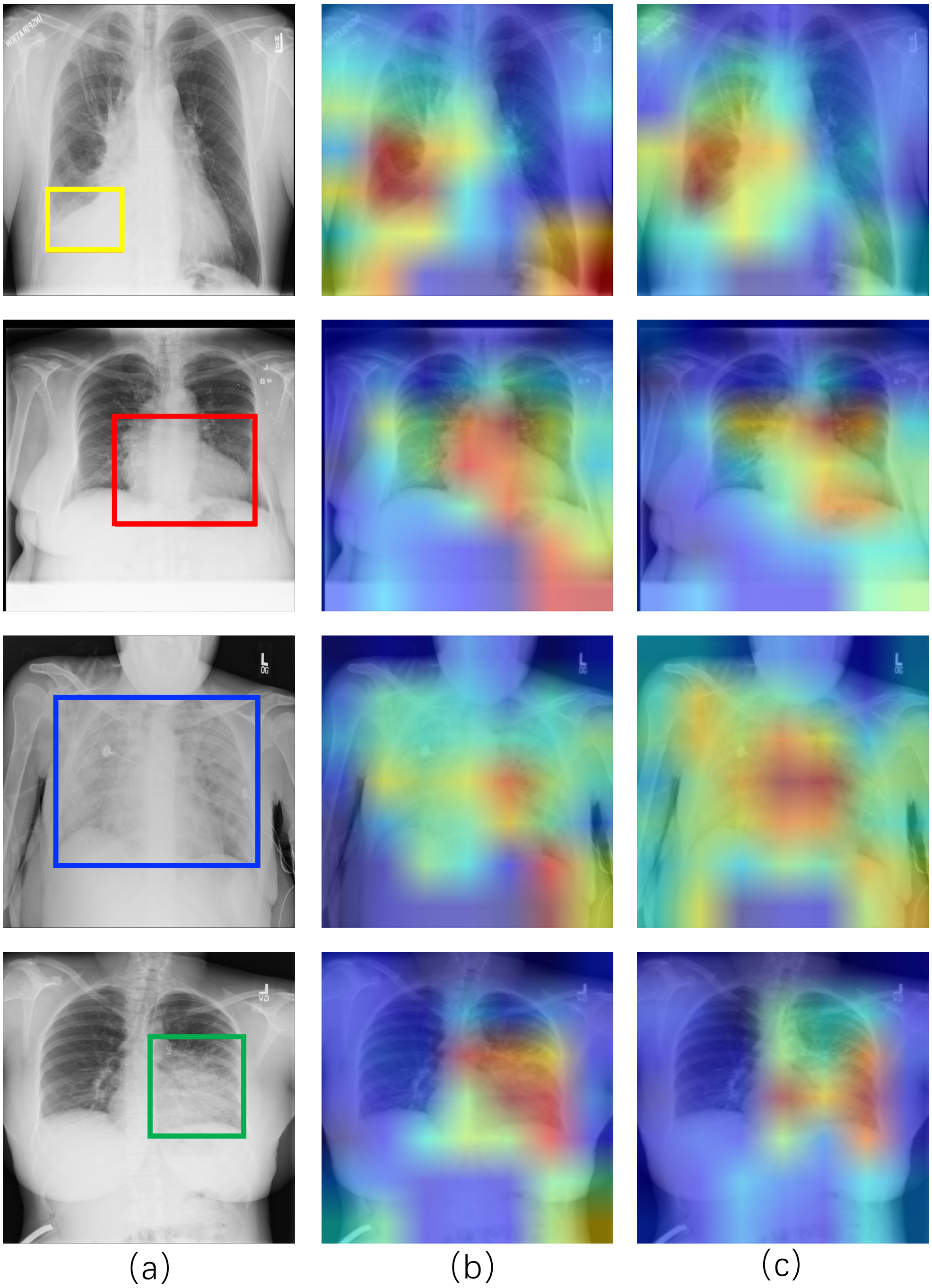


Figure 5.4 : Visualized heatmaps generated by VIB and VSIB. (a) is the input image with the lesion area bounding box annotated by (Wang et al., 2017b). (b) and (c) are the heatmaps generated by the VIB and VSIB constraint, respectively. The large/small response trends to be red/blue in the heatmaps. The larger responses that locate at the position of the corresponding bounding box would be expected.

single channel. We perform the spatial-and-channel encoding as follows: the input feature $\boldsymbol{Z}$ is first fed into three 1×1 convolutional layers whose weights are $\boldsymbol{W}_\theta, \boldsymbol{W}_\phi$, and $\boldsymbol{W}_g$, respectively. To capture the long-range dependencies in spatial and channel dimensions, we then compute the response $\hat{\boldsymbol{Z}} \in \mathbb{R}^{N \times C}$ as

$$vec(\hat{\boldsymbol{Z}}) = f(vec(\boldsymbol{Z}\boldsymbol{W}_\theta), vec(\boldsymbol{Z}\boldsymbol{W}_\phi))vec(\boldsymbol{Z}\boldsymbol{W}_g), \tag{5.9}$$

where $vec(\cdot)$ is a reshape operation that merges the feature dimension from $\mathbb{R}^{N \times C}$ to $\mathbb{R}^Q$ ($Q = N * C$). $f(\cdot, \cdot)$ is a general pairwise function that can differentiate between pairs of same location but at different channels. If two features at different positions are semantic correlated in the feature space, the corresponding $f(\cdot, \cdot)$ between them is learnt and expected to be larger than those not correlated. To simplify the computational complexity, let $\boldsymbol{\theta} = vec(\boldsymbol{Z}\boldsymbol{W}_\theta), \boldsymbol{\phi} = vec(\boldsymbol{Z}\boldsymbol{W}_\phi)$ and $\boldsymbol{g} = vec(\boldsymbol{Z}\boldsymbol{W}_g)$, f is a RBF kernel function that computes a $Q \times Q$ matrix composed by the elements,

$$[f(\boldsymbol{\theta}, \boldsymbol{\phi})]_{i,j} \approx \sum\nolimits_{p=0}^{P} \alpha_p^2 (\theta_i \phi_j)^p, \tag{5.10}$$

which can be approximated by Taylor series up to certain order P. P is set to 3 in our experiment. By introducing two matrices $\boldsymbol{\Theta} = [\alpha_0 \boldsymbol{\theta}^0, \ldots, \alpha_P \boldsymbol{\theta}^P]$ and $\boldsymbol{\Phi} = [\alpha_0 \boldsymbol{\phi}^0, \ldots, \alpha_P \boldsymbol{\phi}^P]$, we can approximate Eq. 5.9 via a trilinear equation,

$$vec(\hat{\boldsymbol{Z}}) \approx \boldsymbol{\Theta}\boldsymbol{\Phi}^\top \boldsymbol{g}. \tag{5.11}$$

The channel grouping idea is then applied to divide the transformed features along the channel dimension into G groups. Each group is approximated according to Eq. 5.11. Then, we wrap the Eq. 5.11 in an identity mapping of the input:

$$\tilde{\boldsymbol{Z}}_{sce} = concat(BN(\hat{\boldsymbol{Z}}\boldsymbol{W}_z)) + \boldsymbol{Z}, \tag{5.12}$$

where $\boldsymbol{W}_z$ is the weights of a 1×1 convolutional layer followed by a Batch Normalization layer with parameters $\boldsymbol{W}_b$. We denote the parameters in *Feature Integrator* module as $\boldsymbol{W}_{sce} = [\boldsymbol{W}_\theta, \boldsymbol{W}_\phi, \boldsymbol{W}_g, \boldsymbol{W}_z, \boldsymbol{W}_b]$.

Finally, we concatenate the features $[\hat{\boldsymbol{Z}}_{sce}, \tilde{\boldsymbol{Z}}_{vsib}]$ as the input of decoder D (with parameters $\boldsymbol{W}_D$). To summarize, in *ConsultNet*, $\boldsymbol{W}$ is the set of learnable parameters, where $\boldsymbol{W} = [\boldsymbol{W}_E, \boldsymbol{W}_{sce}, \boldsymbol{W}_{vsib}, \boldsymbol{W}_D]$.

Note that we assume that multiple diseases exist in the chest X-ray images in previous discussions. While for the situations of only one kind of disease or normal images, the SCE module is also proper because it computes the latent semantic dependencies in the feature space. If there is only one disease, the encoded semantic correlations between the lesion area and other healthy regions would be hindered to make the learned features distinctive for classification. Besides, for a normal image, the semantic correlations among the disease features are relative weaker than those of with multiple diseases.

5.2.4 Optimization with Pairwise Confusion

Chest X-ray image classification task suffers from the large inter-class similarity and a very small number of samples for some certain pathologies. The CXR images do not accurately represent the complete variation because they are visually similar to each other, especially for some patients with two or more pathologies. This similarity can result in overfitting when training CNNs with a large number of parameters, not mention to the categories with a small number of samples. In the ChestX-ray14 dataset, there are only 227 “Hernia” positive samples while the number of all samples is over 100,000. Some loss functions, *e.g.*, triple loss using in the field of object tracking (Dong and Shen, 2018; Dong et al., 2019), could enforce the network learning powerful features by positive-negative sampling. In this work, one sample with category “c_1” is selected as a positive sample for a specific anchor, while it could be a negative for category “c_2” for the same anchor. If using triplet loss, we must treat each category separately, and each category should be assigned an independent classifier. This is not always consistent with the intrinsic relationships among the multiple diseases. Obviously, this ambiguity would also make the network confusing and thus not easy to converge.

To address the above issue, we propose to regularize the *ConsultNet* with a pairwise confusion (PC) strategy (Dubey et al., 2018). The pairwise confusion is introduced in output logits during *ConsultNet* training. It forces the network to learn slightly less discriminative features, thereby preventing it from overfitting to the sample-specific features. Specifically, we aim to confuse the network, by minimizing the distance between the predicted probability distributions for random pairs of samples from the training set. For a pair of samples $\boldsymbol{x}_u, \boldsymbol{x}_v$, the pairwise confusion loss is measured by the Euclidean distance as:

$$\begin{aligned} &\mathcal{L}_{cnf}(p(\hat{\boldsymbol{y}}_u|\boldsymbol{x}_u), p(\hat{\boldsymbol{y}}_v|\boldsymbol{x}_v); \boldsymbol{W}) \\ &= \sum_{c=1}^{\mathcal{C}} (p(\hat{y}_c|\boldsymbol{x}_u) - p(\hat{y}_c|\boldsymbol{x}_v))^2 \\ &= \| \ p(\hat{\boldsymbol{y}}_u|\boldsymbol{x}_u) - p(\hat{\boldsymbol{y}}_v|\boldsymbol{x}_v) \ \|_2^2 \ . \end{aligned} \tag{5.13}$$

For each image, there are N-1 choices for the other image to compute PC loss, giving us a total of $N(N\text{-}1)/2$ possible pairs. N is the number of samples. But in practice, the convergence is achieved after only a fraction of all the possible pairs are observed. Thus, we only evaluate a pairwise confusion loss term on the corresponding pairs of samples across an incoming batch. Pairwise confusion is simple to implement and has no added overhead in training or testing time.

We summarize the whole framework and optimize it by measuring the classification loss and the corresponding pairwise confusion loss of pairs of samples. For each constructed pair, a pair of confusion loss is calculated, together with classification loss for each sample in the constructed pair. We rewrite the total loss for the pair of $\boldsymbol{x}_u, \boldsymbol{x}_v$

$$\begin{aligned} \mathcal{L}(\boldsymbol{W}) = \mathcal{L}_{cls}(\boldsymbol{x}_u, \boldsymbol{y}_u) + \mathcal{L}_{cls}(\boldsymbol{x}_v, \boldsymbol{y}_v) + \\ \gamma\delta(\boldsymbol{y}_u, \boldsymbol{y}_v)\mathcal{L}_{cnf}(\boldsymbol{x}_u, \boldsymbol{x}_v, \boldsymbol{y}_u, \boldsymbol{y}_v) \end{aligned} \tag{5.14}$$

where $\delta(\boldsymbol{y}_u, \boldsymbol{y}_v) = 1$ when $\boldsymbol{y}_u \neq \boldsymbol{y}_v$, and 0 otherwise. γ is a hyper-parameter.

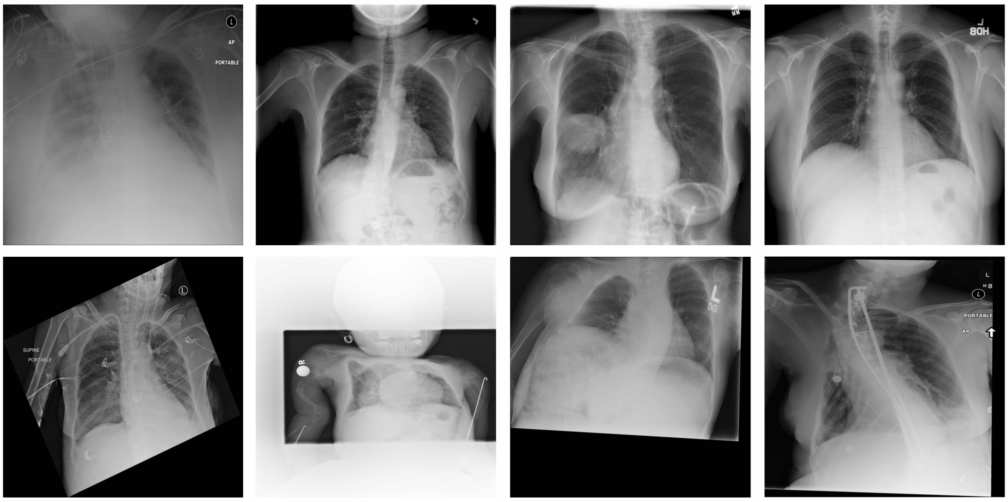


Figure 5.5 : Examples in the ChestX-ray14 dataset. The second row shows some cases captured with abnormal conditions, which introduce noises at the edges of images.

5.3 Experiment

This section evaluates the performance of the proposed *ConsultNet*. We first introduce the experiment datasets and implementation details in Sec. 5.3.1 and Sec. 5.3.2, respectively. Sec. 5.3.3 compares the performance of *ConsultNet* with the state-of-the-art methods. Then we analyze the effectiveness of each component of *ConsultNet* in Sec. 5.3.4.

5.3.1 Datasets

We evaluate our method on the ChestX-ray14 (Wang et al., 2017b) and CheXpert (Irvin et al., 2019) datasets. The AUC (area under the receiver operating characteristic curve) score of each pathology and the average AUC score over all pathologies are reported, respectively.

ChestX-ray14 (Wang et al., 2017b) consists of 112,120 frontal-view X-ray images of 30,805 unique patients. 51,708 images of them are labeled with up to 14 pathologies, while the others are labeled as “No Finding”. Fig. 5.5 presents some examples in ChestX-ray14. In our experiment, we utilize the dataset split provided by (Wang et al., 2017b). It is randomly shuffled the entire dataset into three subgroups on the patient level (86,524 images (80%) for training and validation, 25596 images (20%) for testing). All images from the same patient will only appear in one of the three sets. We report the 14 thoracic disease recognition performance on the published testing set.

CheXpert (Irvin et al., 2019) is a large scale dataset for chest X-rays released by Stanford University. It contains 224,316 chest radiographs of 65,240 patients. 14 observations are labeled in radiology reports, capturing uncertainties inherent in radiography interpretation. We evaluate the *ConsultNet* and compare the performance on the validation set splited by (Irvin et al., 2019). Same with (Irvin et al., 2019), We also evaluate the *ConsultNet* on the 5 competitive pathologies ("Atelectasis", "Cardiomegaly", "Consolidation", "Edema" and "Pleural Effusion").

5.3.2 Implementation Details

We use PyTorch for implementation. DenseNet-121 (Huang et al., 2017) pretrained on the ImageNet (Deng et al., 2009) is used as the backbone of *ConsultNet*. For training, we perform data augmentation by resizing the original images to 256×256, randomly cropping to 224×224, and randomly horizontal flipping. The ImageNet mean value is subtracted from the image. We optimize the network by SGD with a mini-batch size of 64 and train 50 epochs. The learning rate starts from 0.01 and is divided by 10 after 20 epochs. We use a weight decay of 0.0001 and a momentum of 0.9. We empirically set the hyper-parameter $\beta = 1e\text{-}6$, $I_c = 200$ as referred in (Alemi et al., 2017; Peng et al., 2019). γ is set to 0.001. During testing, the image is also resized to 256×256, and then cropped to 224×224.

Table 5.1 : Comparison results of various methods on ChestX-ray14.

Method	CNN	ImgSize	Atel	Card	Effu	Infi	Mass	Nodu	Pne1	Pne2	Cons	Edem	Emph	Fibr	PT	Hern	Mean
(Wang et al., 2017b)	R-50	–	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
(Yao et al., 2018)	∗	512	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
(Shen and Gao, 2018)	–	256	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
(Tang et al., 2018)	–	512	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
(Li et al., 2018)	D-121	299	0.728	0.848	0.782	0.645	0.747	0.702	0.632	0.802	0.727	0.823	0.757	0.763	0.735	0.653	0.739
(Guan and Huang, 2020)	D-121	256	0.781	0.883	0.831	0.697	0.830	0.764	0.725	0.866	0.758	0.853	0.911	0.826	**0.780**	0.918	0.816
baseline	D-121	256	0.778	0.870	0.827	0.698	0.824	0.765	0.729	0.857	0.755	0.849	0.911	0.813	0.773	0.827	0.806
ConsultNet	D-121	256	**0.785**	**0.899**	**0.835**	0.699	**0.838**	0.775	0.738	**0.871**	0.763	**0.850**	**0.924**	**0.831**	0.776	**0.922**	**0.822**

[*] The 14 pathologies are Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, Pneumothorax, Consolidation, Edema, Emphysema, Fibrosis, Pleural Thickening and Hernia, respectively. For each column, the best results are highlighted in bold. * represents that the combination of ResNet and DenseNet is used in Yao *et al.* (Yao et al., 2018). – represents the network used in the corresponding reference is not illustrated. we compare the best performance in (Wang et al., 2017b; Shen and Gao, 2018; Tang et al., 2018; Guan and Huang, 2020).

Table 5.2 : Comparison results of various methods on the CheXpert dataset.

Method	Models	Policy	ImgSize	Atelectasis	Cardiomegaly	Consolidation	Edema	Pleural Effusion	Mean
(Irvin et al., 2019)	ensemble	Zeros	320	**0.811**	0.840	0.932	**0.929**	**0.931**	**0.889**
(Pham et al., 2019)	single	Zeros	256	0.806	0.833	0.929	0.933	0.921	0.884
Baseline	single	Zeros	256	0.799	0.832	0.927	0.897	0.923	0.875
ConsultNet	single	Zeros	256	0.804	**0.874**	**0.940**	0.894	0.923	**0.889**
(Irvin et al., 2019)	ensemble	Ones	320	**0.858**	0.832	0.899	**0.941**	**0.934**	0.893
(Pham et al., 2019)	single	Ones	256	0.825	0.855	0.937	0.930	0.923	0.894
Baseline	single	Ones	256	0.772	0.845	**0.940**	0.908	0.925	0.878
ConsultNet	single	Ones	256	0.847	**0.868**	0.923	0.924	0.926	**0.898**

[*] DenseNet-121 is utilized as our baseline on the CheXpert dataset. “Zeros” and “Ones” are the different setting for the uncertainy label. The best performance for each pathology is in bold.

5.3.3 Comparative Studies

The DenseNet-121 (Huang et al., 2017) is adopted as our baseline on both ChestX-ray14 and CheXpert dataset. The architecture configuration is the same as (Huang et al., 2017) except for replacing the last global average pooling with global max pooling, the original classifier with a 14- and 5-dimensional fully connected layer for ChestXray14 and CheXpert, respectively. The AUC scores are reported in Tab. 5.1 and Tab. 5.2.

Results on ChestX-ray14. We first report the performance of the baseline. On the ChestXray14 dataset, the AUC score of each pathology is summarized in Tab. 5.1. The average AUC score of our baseline arrives at 0.806 across the 14 thorax diseases. It is competitive or even better than the previous works except for (Guan and Huang, 2020) in Tab. 5.1. For some of the 14 pathologies, *e.g.*, "Nodule" (0.765 vs. 0.764) and "Pneumonia" (0.729 vs. 0.725), the performance of our baseline is very close to (Guan and Huang, 2020). The AUC of "Infiltration" is little higher (0.698 vs. 0.693) than (Guan and Huang, 2020). Moreover, we observe that "Infiltration" has lower recognition accuracy among 14 pathologies. It is because the diagnosis of "Infiltration" mainly relies on the subtle texture change among the lung area, which is challenging to recognize. For some categories with a small number of samples, *e.g.*, "Pleural Thickening" or "Edema", there is still a large gap between our baseline network and (Guan and Huang, 2020).

In Tab. 5.1, the AUC scores of 14 pathologies are presented, indicating the effectiveness of the proposed method. The average AUC score has 0.822, which is higher than the baseline 0.16. By introducing SCE, the diseases highly correlating with other ones have a significant performance improvement, for example, 2% for "Atelectasis" and 1.4% for "Consolidation". Without surprise, the diseases without strong correlations among other ones do not receive such significant benefits after introducing SCE (only 0.2% for "Edema"). Except "Infiltration", our method obviously improves the baseline over other pathologies. The AUC scores of pathology like "Atelectasis", "Cardiomegaly", "Fibrosis", "Mass", or "Hernia" consistently exceed the baseline about 2%. Particularly, the AUC of "Emphyseme" improves about 3% from 0.893 to 0.924. The improvement on the other pathologies, such as "Edema", and "Pleural Thickening", is not so significant. We compare our results with the state-of-the-art methods (Wang et al., 2017b; Li et al., 2018; Tang et al., 2018; Shen and Gao, 2018; Guan and Huang, 2020). The best performance of each pathology is shown in bold. Compared with other methods, *ConsultNet* shows its priority in most of the 14 pathologies. For "Fibrosis", "Nodule", "Cardiomegaly", and "Emphysema", our method yields better performance (over 1%) compared with (Guan and Huang, 2020). We are slightly lower than (Guan and Huang, 2020) on "Edema" and "Pleural Thickening". The same situation happens for "Consolidation" compared with (Shen and Gao, 2018). It is worth to explore that (Shen and Gao, 2018) produces a very surprising performance for "Infiltration" and "Pneumonia". We conjecture that the reason might be the utilization of a shallow network in (Shen and Gao, 2018). The shallow neural network is more suitable for capturing the low-level information, which is sufficient to recognize "Infiltration" and "Pneumonia" from images.

Results on CheXpert. We report the performance of the *ConsultNet* on the CheXpert dataset in this section. On the CheXpert dataset, due to the setting of uncertain training labels, we need to explore different approaches to use the uncertainty labels during the model training. To evaluate the effectiveness of *ConsultNet*, we utilize two common uncertain label policies in multi-label classification (Kolesov et al., 2014): 1) replacing all the uncertain labels by the "zeros"; 2) replacing all the uncertain labels by "ones". The AUC scores are presented in Tab. 5.2. Same as the experiment on ChestX-ray14, DenseNet-121 is used as the baseline. We present the performance obtained by a single model rather than the performance of Irvin *et al.*(Irvin et al., 2019), which is from the ensemble of 30 models.

In our baseline, the average AUC scores over five pathologies achieve 0.875 and 0.878 for "zeros" and "ones" policy, respectively. With *ConsultNet*, both of them are improved over 1%. Particularly, when the uncertain labels are set to “ones”, the average AUC score surpasses the corresponding baseline about 2%. Here, we notice that *ConsultNet* shows its superiority for some pathologies in different uncertain label policy. The performance of “Cardiomegaly” and “Consolidation” are significantly improved (0.832 vs 0.874, 0.927 vs 0.940) when the uncertain labels are set to “zeros”. While setting to “ones”, *ConsultNet* surpasses the baseline a large margin on the AUC scores of “Atelectasis” and “Cardiomegaly” (0.772 vs 0.847, 0.845 vs 0.868). We also notice that the performance of “Consolidation” reduces and does not present the same trend like with “zeros” policy. We analyze the main reason is that it is unreasonable to set the specific uncertainty of “Consolidation” to “ones”. Apart from this, the performance of “Edema” and “Pleural Effusion” is not improved too much by *ConsultNet*. But as refer to the label policies, we conclude that “ones” is much more proper than “zeros” because of the performance. Note that more proper uncertainty strategy or learning technology could be explored to achieve better performance. But we do not consider that because it is out of the range of this work.

We compare the proposed *ConsultNet* with the state-of-the-art methods. The best performance of each uncertain label setting for each pathology is shown in bold in Tab. 5.2. We test the *ConsultNet* with only a single model, but not with the ensemble of tens of models. The average AUC scores of *ConsultNet* achieve to 0.889 and 0.898 for “Zeros” and “Ones” policy, respectively. Compared with other methods, the performance is comparative or even better on both “Zeros” and “Ones” policy. For “Cardiomegaly” and “Consolidation”, the AUC scores of *ConsultNet* exceed the other methods a large margin. For example, the performance of “Cardiomegaly” is improved about 3% and 2% for “Zeros” and “Ones”, respectively. For the other three pathologies, *ConsultNet* also performs not far-off compared with the methods tested with single model. It is merely inferior to (Irvin et al., 2019), whose performance is obtained by the ensemble of 30 models.

Further Analysis for Different Image Resolutions. Considering that the image size used in previous experiments may not be appropriate for the clinical practice usage, we conduct experiments on the large image resolutions and discuss the effect of image size on the final performance of *ConsultNet*. Training *ConsultNet* with different image resolutions would not affect the training processes, but the final performance. First, due to a global max pooling layer is following the feature extraction backbone, the mid-level features with different dimensions in spatial would be unified into same dimension finally. Therefore, the training or testing processes of *ConsultNet* would not be changed. Second, high-resolution images could provide more critical details of lesion area and thus training with them achieves higher performance compared with low-resolution images.

In this experiment, we set the training image size to 512×512 and 1024×1024. The other experimental settings are the same as those in the previous experiment, but adaptive adjusting the training batchsize based on the GPU resources used. The training batchsize for 512×512 and 1024×1024 are 32 and 16, respectively. Table 5.3 and Table 5.4 present the AUC scores of each pathology and the average AUC score over all pathologies on the ChestX-ray and CheXpert dataset, respectively. All the average AUC scores are largely improved by increasing the training image size. On the ChestX-ray14 dataset, the average AUC scores of *ConsultNet* and the baseline methods training with 512×512 images surpass that with 256×256 nearly 1%. While training with larger images (1024×1024), the average AUC scores of the baseline and *ConsultNet* have 0.820 and 0.841, which are improved further. Training with 1024×1024 images, the AUC scores of most of the 14 pathologies exceed those of lower resolution images. While for some pathologies (*e.g.*, "Infiltration", "Nodule", and "PT"), their AUC scores for 512×512 are better than the other two resolutions. On the CheXpert dataset, there are over 1% improvement when training the baseline and *ConsultNet* methods with 512×512 images compared with 256×256 images for both "Zeros" and "Ones" policy. The performance of training with 1024×1024 images achieves 0.913 and 0.921, which are the state of the art. Besides, their average AUC scores exceed the corresponding baseline method about 2% (0.913 vs. 0.893, 0.921 vs. 0.904).

Generally, training with low-resolution chest X-ray images achieves lower performance compared with high-resolution images. This might be caused by critical information losing (*e.g.*, image details related with lesion area) produced by down-sampling or data augmentation techniques. We empirically conclude the above observation under the condition of training *ConsultNet* based on the fixed hyperparameters of SGD. While training with other hyperparameters, *e.g.*, using a grid search to find optimal hyperparameters (learning rate, batch size, *etc.*) in (Tang et al., 2020), such conclusion might be different. (Tang et al., 2020) empirically finds that the performance is less impacted by the input image size in the normal vs. abnormal binary classification case.

Computational Consumption Analysis. Except for the training images, the computational consumption is also a factor that should be considered for clinical usage. We mainly concentrate on the model parameters and input image size that affect the GPU consumption of *ConsultNet*. For a certain backbone, *e.g.*, DenseNet-121 used in *ConsultNet*, the scale of model parameters of *ConsultNet* increases from 7.98M to 11.82M. Taking 256×256 input image as an example, the FLOPs (floating point operations) for *ConsultNet* is 2.96G MACs (Multiplication and Accumulation), which increases not two much (2.88G MACs for the baseline method). In practice, training *ConsultNet* with 256×256 images on a TiTAN XP GPU of 12 GB memory, it costs nearly 0.35 second per image. Large resolution images would cost much more time. The FLOPs increase about 0.32G (11.85G vs. 11.53G) MACs and 1.28G (47.39G vs. 46.11G) MACs for 512×512 and 1024×1024, respectively. Besides, the GPU consumption is extremely different when loading the training data with different input image size or batchsize. In our experiment, when training *ConsultNet* with 64 images (256×256) in a mini-batch, it costs about 9GB GPU memory. While the input is set to 16 images (1024×1024) in a mini-batch, nearly 42 GB GPU memory is required. Training with larger resolution images achieves better performance in the above experiments, but the amount of GPU memory used also increases accordingly.

Table 5.3 : Comparison of various image resolutions on the ChestX-ray Dataset.

Method	ImgSize	Atel	Card	Effu	Infi	Mass	Nodu	Pne1	Pne2	Cons	Edem	Emph	Fibr	PT	Hern	Mean
Baseline	256	0.778	0.870	0.827	0.698	0.824	0.765	0.729	0.857	0.755	0.849	0.911	0.813	0.773	0.827	0.806
ConsultNet	256	0.785	0.899	0.835	0.699	0.838	0.775	0.738	0.871	0.763	0.850	0.924	0.831	0.776	0.922	0.822
Baseline	512	0.781	0.884	0.832	0.699	0.827	0.777	0.725	0.862	0.753	0.851	0.910	0.818	0.773	0.908	0.814
ConsultNet	512	0.797	0.909	0.848	**0.709**	0.848	**0.789**	0.740	0.874	0.779	0.858	0.929	0.834	**0.796**	0.928	0.831
Baseline	1024	0.786	0.898	0.835	0.695	0.844	0.770	0.735	0.866	0.757	0.855	0.918	0.837	0.774	0.916	0.820
ConsultNet	1024	**0.809**	**0.911**	**0.851**	0.706	**0.861**	0.776	**0.777**	**0.900**	**0.790**	**0.864**	**0.939**	**0.857**	0.787	**0.948**	**0.841**

The 14 pathologies are same as those in Table 5.1.

Table 5.4 : Comparison of various image resolutions on the CheXpert Dataset.

Policy	Method	ImgSize	Atelectasis	Cardiomegaly	Consolidation	Edema	Pleural Effusion	Mean
Zeros	Baseline	256	0.799	0.832	0.927	0.897	0.923	0.875
	ConsultNet	256	0.804	0.874	0.940	0.894	0.923	0.889
	Baseline	512	0.830	0.845	0.921	0.898	0.930	0.885
	ConsultNet	512	0.836	0.880	0.941	0.923	0.936	0.903
	Baseline	1024	0.839	0.850	0.933	0.909	0.933	0.893
	ConsultNet	1024	**0.856**	**0.887**	**0.946**	**0.933**	**0.943**	**0.913**
Ones	Baseline	256	0.772	0.845	0.940	0.908	0.925	0.878
	ConsultNet	256	0.847	0.868	0.923	0.924	0.926	0.898
	Baseline	512	0.836	0.846	0.936	0.929	0.926	0.895
	ConsultNet	512	0.849	0.868	0.939	0.948	0.946	0.910
	Baseline	1024	0.850	0.862	0.938	0.942	0.929	0.904
	ConsultNet	1024	**0.866**	**0.879**	**0.947**	**0.953**	**0.958**	**0.921**

Table 5.5 : Ablation study on ChestX-ray14. The average AUC scores over 14 patholgies are reported.

VSIB	SCE	PC	Mean
			0.8059
√			0.8166
	√		0.8158
√	√		0.8206
√	√	√	**0.8220**

5.3.4 Effectiveness of *ConsultNet*

This section evaluates the effectiveness of each component of *ConsultNet* on the ChestX-ray14 dataset. We first present the ablation studies and visualize the learned heatmaps by each module in *ConsultNet*. Then we show the superiority of the proposed VSIB constraint qualitatively and quantitatively. We visualize the learned heatmaps constrained by the proposed VSIB and the original VIB, and compare the performance produced by them.

Ablation Study. We remove one module in VSIB/SCE/PC but activate the left ones at a time. Tab. 5.5 presents the average AUC score over 14 pathologies on the ChestX-ray4 dataset. Totally, *ConsultNet* surpasses the baseline about 1.6%. Introducing VSIB or SCE, the average AUC exceeds the baseline about 1%. By combing them together, the performance achieves 0.8206 and is further improved. The performance is slightly enhanced after introducing PC. The PC loss could induce the risk of deep network overfitting on the high appearance similarity of chest X-ray images, especially on the small number of positive samples. The AUC scores are improved obviously for some pathologies with a small number of samples, *e.g.*, the AUC of “Hernia” is improved over 1% (from 0.911 to 0.922). For other pathologies, the improvement is not so significant, *e.g.*,

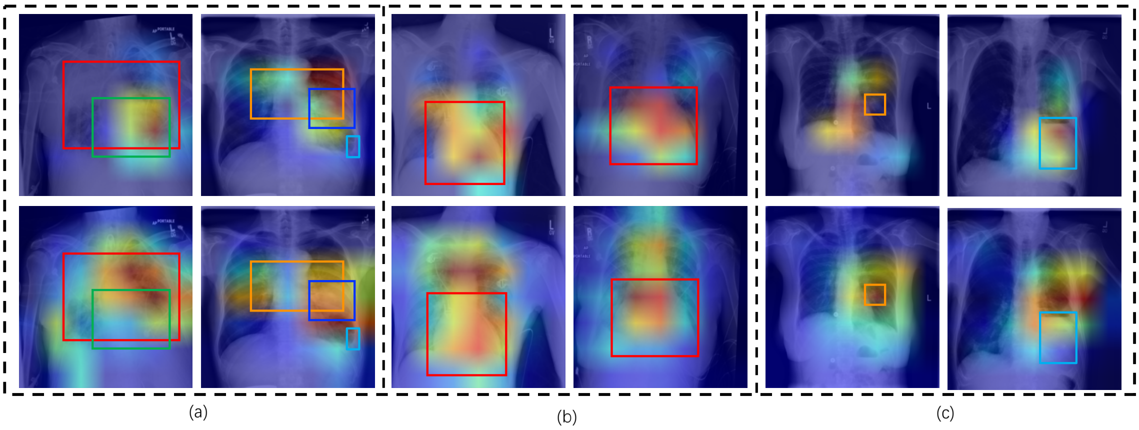


Figure 5.6 : Examples of heatmaps generated from VSIB (the first row) and SCE (the second row). (a) The VIB and SCE collaboratively focus on different lesion areas. (b) The VSIB module learns much more accurate lesion positions while SCE misjudges some healthy tissues. (c) The VSIB module misses recognizing the disease existing in the image while the SCE module complements the VSIB module in the first column. While in the second column, VSIB module successes to localize the accurate disease region, but the region localized by SCE module drifts out of the truly lesion area. The different color of bounding box presents the different pathology existing in the image. Same as Fig. 5.4, the position of bounding box is the ground truth provided in (Wang et al., 2017b). The larger responses locate at the position of the corresponding bounding box are expected.

from 0.780 to 0.785 for “Atelectasis”. This discrepancy may be caused by the number of positive samples (*e.g.*, only 227 “Hernia” samples while 7323 “Atelectasis” samples).

Visualization. We visualize the feature heatmaps of VSIB and SCE modules for some examples on the ChestX-ray14 dataset. The heatmaps are shown in Fig. 5.6. We concentrate on CXR image classification and thus do not use a threshold to localize the heatmap or compute the IoU. Each heatmap is generated by calculating the maximized absolute activation value along the feature channel. The heatmaps are resized to the same spatial dimension as the input images. As shown in Fig. 5.6 (a), the VSIB and SCE modules collaboratively focus on different lesion areas. The SCE module fires at the regions of both

Table 5.6 : The localization accuracy between VSIB and VIB.

T(IoU)	Model	Atel	Card	Effu	Infi	Mass	Nodu	Pne1	Pne2	Mean
0.1	VIB	0.53	0.83	0.62	0.62	0.58	0.51	0.54	0.50	0.54
	VSIB	0.57	0.88	0.68	0.68	0.67	0.59	0.62	0.59	0.57
0.3	VIB	0.21	0.61	0.33	0.33	0.29	0.26	0.27	0.25	0.32
	VSIB	0.25	0.62	0.36	0.36	0.35	0.32	0.33	0.31	0.33
0.5	VIB	0.11	0.34	0.10	0.10	0.09	0.08	0.08	0.07	0.16
	VSIB	0.17	0.41	0.17	0.17	0.17	0.16	0.16	0.15	0.20

* The 8 pathologies are presented as same as that in Tab. 5.1.

the lesion area and some healthy tissues in Fig. 5.6 (b), but VSIB localizes much more accurate positions. Fig. 5.6 (c) presents examples of two modules work collaboratively. When one module fails to recognize a disease correctly, the other module could provide extra supplementary information for diagnosis. The VSIB and SCE modules work collaboratively to make an accurate diagnosis in the proposed framework. They learn the discriminative features for classification from different views. The VSIB module filters the critical features for classification, and the SCE module encodes the semantic dependencies in the feature space. Fusing the multi-view features could provide more information and benefit to recognizing diseases. Specifically, the VSIB module misses recognizing the disease while the SCE module complements the VSIB module in the first column of Fig. 5.6 (c). While in the second column of Fig. 5.6 (c), VSIB module successes to localize the accurate disease region, but the region localized by SCE module drifts out of the truly lesion area.

Comparing VSIB with vanilla VIB. We demonstrate the superiority of the VSIB constraint in the *ConsultNet* from two aspects in this section. We first compare the classification performance between the proposed VSIB and the vanilla VIB constraint. The visualized heatmaps generated by VIB and VSIB are shown in Fig. 5.4. The given bounding boxes on the images show the ground truth of lesion areas. The VSIB module localizes much more centralized and accurate disease regions compared with the VIB module. With DenseNet-121 as backbone, the average AUC of VSIB achieves 0.817 which surpasses VIB by 0.5% (0.812). There are about 1% improvement for some diseases, *e.g.*, "Consolidation" (0.9%) or "Hernia"(1.1%). We believe this is because of two reasons: 1) VSIB explicitly adopts a selective information constraint M on the KL-loss term in Eq. 5.7 and therefore provides additional guidance when learning disease-specific features; 2) VSIB considers feature discrepancies while VIB fails. Moreover, we also evaluate whether the combination of the VSIB and SCE modules works better than that of VIB and SCE. Combining the VIB and SCE modules, the average AUC score has 0.816 which surpasses the VIB about 0.4%. We could conclude that fusing these features benefits to disease recognition. While comparing with the combination of VSIB and SCE, under the same experimental setting, the average AUC score of combing VIB and SCE is inferior to that of VSIB and SCE (0.821). It illustrates that the proposed VSIB constraint is effectiveness no matter working individually or together with the SCE module.

Besides, we provide the localization accuracy between the VSIB and VIB on the ChestX-ray14 dataset. The ChestX-ray14 dataset (Wang et al., 2017b) provides 984 bounding boxes of 8 pathologies. Intersection over Union (IoU) is computed between the region localized by VSIB (or VIB) and the ground truth. We define an accurate localization by requiring its IoU is greater than a threshold of IoU (T(IoU)). Tab. 5.6 presents the localization accuracy of VIB and VSIB over 8 pathologies with different threshold (T(IoU) = $\{0.1, 0.3, 0.5\}$). The localization accuracy of VSIB obviously surpasses VIB over different threshold of IoU. In VSIB, we focus on learning the disease-critical features for classification, whose purpose is similar to the task of saliency detection (Wang et al., 2019). However, there is no "disease-critical" (lesion area ground-truth) given in discriminative feature learning in this work, which is not the same as saliency detection.

However, once a few annotated ground-truth of lesion area is given, the disease classification would benefit from it. We would like to put it into future work.

5.4 Conclusion

This chapter proposes a *ConsultNet* to learn powerful feature representations for thorax disease classification. The proposed *ConsultNet* aims to address the problems of irrelevant regions influence and existing of multiple diseases in chest X-ray image classification. The *ConsultNet* learns disease-specific features by introducing a novel variational selective information bottleneck constraint and explores the latent semantic dependencies of multiple diseases in the feature space. The proposed two modules collaboratively learn discriminative features for chest X-ray image classification. Moreover, a pairwise confusion regularizing strategy is used to address the sample appearance similarity problem. Experimental results show that the proposed method yields better performance and outperforms the baseline quantitatively.

Chapter 6

Conclusions and Future Work

6.1 Summary of Contributions

In this thesis, we investigate the deep learning methods to chest X-ray image classification. Focus on the characteristics of the special domain of chest X-ray images, we contribute to the community with the following three aspectes.

First, we introduce an attention guided convolution neural network (AG-CNN) to classify the chest X-ray images. An attention guided mask inference based cropping strategy is proposed to localize the discriminative region. AG-CNN benefits from combining the global and critical local cues and improves the performance of chest X-ray image classification..

Second, we exploit the correlations of multiple diseases in chest X-ray images to boost the recognition performance. A category-wise residual attention learning (CRAL) method is proposed to achieve this purpose. CRAL re-encodes the image features by attention scores, which are used to suppress the obstacles of irrelevant classes and strengthen the relevant features at the same time. Experiments on the ChestX-ray14 dataset demonstrate the effectiveness of CRAL.

Last, we design a robust and stable chest X-ray image classification method that is able to 1) automatically focus on the disease-critical regions, which usually are of small sizes; 2) adaptively capture the intrinsic relationships among different disease features and utilize them to boost the multi-label disease recognition rates jointly. A framework named ConsultNet is proposed to learn the discriminative features to achieve those two purposes simultaneously. Extensive experiments are conducted to show the superiority of ConsultNet.

6.2 Future Directions

In future work, we will continue to develop more robust and high-performance methods for thorax disease diagnosis. The following aspects are worth considering.

Learning with noisy labels. The current large-scale chest X-ray image datasets are labeled with natural language processing techniques. Many noisy labels or uncertain data are inevitably introduced. A robust deep learning model needs to be able to tolerate noisy or uncertain data. That is, even incorrect annotations existing, it is expected that the learned model could also learn discriminative features to predicate accurately.

Learning with multi-view data. In clinical practice, CT is another common examination for thorax disease diagnosis. Besides, medical reports could provide valuable information for thorax disease classification and localization. Fusing the data from different views, *e.g.*, medical reports, X-ray images, and CTs, could help improve the performance of thorax disease recognition. Learning with multi-view data could achieve high performance and could also make the computer-aided diagnosis system diversified for thorax disease analysis.

Learning medical report automatic generation. Based on the experience of disease recognition and localization, it is natural to develop an automatic medical report generation system. Medical report generation could further provide auxiliary information for clinical diagnosis, and it could also make the computer-aided diagnosis more intelligent. Medical report generation integrates the technologies from the field of computer vision, natural language processing. This makes it more challenging.

Bibliography

Alemi, A. A., Fischer, I., Dillon, J. V. & Murphy, K., 2017, 'Deep variational information bottleneck', *International Conference on Learning Representations*, . 59, 60, 68

Cai, J., Lu, L., Harrison, A. P., Shi, X., Chen, P. & Yang, L., 2018, 'Iterative attention mining for weakly supervised thoracic disease pattern localization in chest X-rays', *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, pp. 589–598. 7

Chen, B., Li, J., Lu, G., Yu, H. & Zhang, D., 2020a, 'Label co-occurrence learning with graph convolutional networks for multi-label chest x-ray image classification', *IEEE Journal of Biomedical and Health Informatics*. 2, 7, 8

Chen, B., Zhang, Z., Lin, J., Chen, Y. & Lu, G., 2020b, 'Two-stream collaborative network for multi-label chest x-ray image classification with lung segmentation', *Pattern Recognition Letters*, vol. 135, pp. 221–227, <`https://www.sciencedirect.com/science/article/pii/S0167865520301380`>. 6

Chen, H., Miao, S., Xu, D., Hager, G. D. & Harrison, A. P., 2020c, 'Deep hiearchical multi-label classification applied to chest x-ray abnormality taxonomies', *Medical Image Analysis*, vol. 66, p. 101811, <`https://www.sciencedirect.com/science/article/pii/S1361841520301754`>. 5, 8

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. & Fei-Fei, L., 2009, 'Imagenet: A large-scale hierarchical image database', *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 248–255. 1, 11, 68

Dong, X. & Shen, J., 2018, ‘Triplet loss in siamese network for object tracking’, *Proceedings of the European Conference on Computer Vision (ECCV)*, . 65

Dong, X., Shen, J., Wu, D., Guo, K., Jin, X. & Porikli, F., 2019, ‘Quadruplet network with one-shot learning for fast visual object tracking’, *IEEE transactions on Image Procesing*, vol. 28, pp. 3156–3527. 65

Dubey, A., Gupta, O., Guo, P., Raskar, R., Farrell, R. & Naik, N., 2018, ‘Pairwise confusion for fine-grained visual classification’, *ECCV*, pp. 70–86. 56, 66

Gohagan, J., Prorok, P., Hayes, R. & Kramer, B., 2000, ‘The prostate, lung, colorectal and ovarian (plco) cancer screening trial of the national cancer institute: history, organization, and status.’, *Controlled clinical trials*, vol. 21, no. 6, pp. 251S–272S. 6

Gong, X., Xia, X., Zhu, W., Zhang, B., Doermann, D. & Zhuo, L., 2021, ‘Deformable gabor feature networks for biomedical image classification’, *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4004–4012. 5

Guan, Q. & Huang, Y., 2020, ‘Multi-label chest X-ray image classification via category-wise residual attention learning’, *Pattern Recognition Letters*, vol. 130, pp. 259–266. 69, 71, 72

Guan, Q., Huang, Y., Zhong, Z., Zheng, Z., Zheng, L. & Yang, Y., 2020, ‘Thorax disease classification with attention guided convolutional neural network’, *Pattern Recognition Letters*, vol. 131, pp. 38–45. 54

Guendel, S., Grbic, S., Georgescu, B., Liu, S., Maier, A. & Comaniciu, D., 2018, ‘Learning to recognize abnormalities in chest x-rays with location-aware dense networks’, *Iberoamerican Congress on Pattern Recognition*, Springer, pp. 757–765. 2, 6, 46, 47, 48, 54

He, K., Zhang, X., Ren, S. & Sun, J., 2016, 'Deep residual learning for image recognition', *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778. 5, 15, 20, 36, 44

Hermoza, R., Maicas, G., Nascimento, J. C. & Carneiro, G., 2020, 'Region proposals for saliency map refinement for weakly-supervised disease localisation and classification', *International Conference on Medical image computing and computer-assisted intervention.*, . 7

Hochreiter, S. & Schmidhuber, J., 1997, 'Long short-term memory', *Neural computation*, vol. 9, no. 8, pp. 1735–1780. 8, 34

Hu, J., Shen, L. & Sun, G., 2018, 'Squeeze-and-excitation networks', *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141. 5

Hu, Y., Zhang, Y., Zhang, T., Gao, S. & Fan, W., 2020, 'Label generation network based on self-selected historical information for multiple disease classification on chest radiography', *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1015–1019. 8

Huang, G., Liu, Z., Weinberger, K. Q. & van der Maaten, L., 2017, 'Densely connected convolutional networks', *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708. 5, 20, 36, 44, 59, 60, 68, 71

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K. et al., 2019, 'Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison', *Association for the Advancement of Artificial Intelligence (AAAI)*, pp. 590–597. 5, 6, 67, 68, 70, 72, 73

Kim, M., Park, J., Na, S., Park, C. M. & Yoo, D., 2020, 'Learning visual context by comparison', *Proceedings of the European Conference on Computer Vision (ECCV)*, . 7

Kingma, D. P. & Welling, M., 2013, 'Auto-encoding variational bayes', *arXiv preprint arXiv:1312.6114*. 60

Kolesov, A., Kamyshenkov, D., Litovchenko, M., Smekalova, E., Golovizin, A. & Zhavoronkov, A., 2014, 'On multilabel classification methods of incompletely labeled biomedical text data', *Computational and mathematical methods in medicine*, vol. 2014. 72

Kumar, P., Grewal, M. & Srivastava, M. M., 2018, 'Boosted cascaded convnets for multilabel classification of thoracic diseases in chest radiographs', *International Conference Image Analysis and Recognition*, Springer, pp. 546–552. 8, 11, 21, 22, 28, 29, 34, 46, 55, 62

Li, Z., Wang, C., Han, M., Xue, Y., Wei, W., Li, L.-J. & Li, F.-F., 2018, 'Thoracic disease identification and localization with limited supervision', *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8290–8299. 23, 24, 25, 46, 47, 48, 54, 69, 72

Liang, X., Peng, C., Qiu, B. & Li, B., 2019, 'Dense networks with relative location awareness for thorax disease identification', *Medical Physics*, vol. 46, no. 5, pp. 2064–2073. 6, 7

Liu, H., Wang, L., Nan, Y., Jin, F., Wang, Q. & Pu, J., 2019, 'Sdfn: Segmentation-based deep fusion network for thoracic disease classification in chest x-ray images', *Computerized Medical Imaging and Graphics*, vol. 75, pp. 66–73, <https://www.sciencedirect.com/science/article/pii/S0895611118306177>. 6

Mo, S. & Cai, M., 2019, 'Deep learning based multi-label chest x-ray classification with entropy weighting loss', *2019 12th International Symposium on Computational Intelligence and Design (ISCID)*, , vol. 2pp. 124–127. 8

Olaf, R. & Fisher, T., Philippand Brox, 2015, 'U-net: Convolutional networks for biomedical image segmentation.', *International Conference on Medical image computing and computer-assisted intervention.*, . 6

Paszke, A., Gross, S., Chintala, S. & Chanan, G., 2017, 'Pytorch', . 44

Peng, X. B., Kanazawa, A., Toyer, S., Abbeel, P. & Levine, S., 2019, 'Variational discriminator bottleneck: Improving imitation learning, inverse rl, and gans by constraining information flow', *International Conference on Learning Representations*. 68

Pham, H. H., Le, T. T., Tran, D. Q., Ngo, D. T. & Nguyen, H. Q., 2019, 'Interpreting chest x-rays via cnns that exploit disease dependencies and uncertainty labels', *arXiv preprint arXiv:1911.06475*. 70

Rajpurkar, P., Irvin, J. & et. al., 2017, 'Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning', *arXiv preprint arXiv:1711.05225*. 5, 20, 21, 28, 29, 46

Seibold1, C., Kleesiek, J., Schlemmer, H.-P. & Stiefelhagen, R., 2020, 'Self-guided multiple instance learning for weakly supervised disease classification and localization in chest radiographs', *Proceedings of the Asial Conference on Computer Vision (ACCV)*, . 7

Shen, Y. & Gao, M., 2018, 'Dynamic routing on deep neural network for thoracic disease classification and sensitive area localization', *International Workshop on Machine Learning in Medical Imaging*, pp. 389–397. 5, 46, 48, 69, 72

Shiraishi, J., Katsuragawa, S., Ikezoe, J., Matsumoto, T., Kobayashi, T., Komatsu, K.-i., Matsui, M., Fujita, H., Kodera, Y. & Doi, K., 2000, 'Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules', *American Journal of Roentgenology*, vol. 174, no. 1, pp. 71–74. 6

Simonyan, K. & Zisserman, A., 2014, ‘Very deep convolutional networks for large-scale image recognition’, *arXiv preprint arXiv:1409.1556*. 5

Szegedy, C., Liu, W. & et. al., 2015, ‘Going deeper with convolutions’, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9. 5

Taghanaki, A. I. & Havaei, M., 2019, ‘Infomask: Masked variational latent representation to localize chest disease’, *MICCAI*, . 9

Tang, Y., Wang, X., Harrison, A. P., Lu, L., Xiao, J. & Summers, R. M., 2018, ‘Attention-guided curriculum learning for weakly supervised classification and localization of thoracic diseases on chest radiographs’, *International Workshop on Machine Learning in Medical Imaging*, Springer, pp. 249–258. 46, 48, 54, 55, 69, 72

Tang, Y.-X., Tang, Y.-B., Peng, Y., Yan, K., Bagheri, M., Redd, B. A., Brandon, C. J., Lu, Z., Han, M., Xiao, J. et al., 2020, ‘Automated abnormality classification of chest radiographs using deep convolutional neural networks’, *NPJ Digital Medicine*, vol. 3, no. 1, pp. 1–8. 75

TISHBY, N., 1999, ‘The information bottleneck method’, *Proc. 37th Annual Allerton Conference on Communications, Control and Computing, 1999*, pp. 368–377. 9

Tishby, N. & Zaslavsky, N., 2015, ‘Deep learning and the information bottleneck principle’, *IEEE Information Theory Workshop*, pp. 1–5. 59, 60

Van Ginneken, B., Stegmann, M. B. & Loog, M., 2006, ‘Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database’, *Medical Image Analysis*, vol. 10, no. 1, pp. 19–40. 6

Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X. & Tang, X., 2017a, ‘Residual attention network for image classification’, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3156–3164. 36, 40

Wang, K., Zhang, X., Huang, S., Chen, F., Zhang, X. & Huangfu, L., 2020, 'Learning to recognize thoracic disease in chest x-rays with knowledge-guided deep zoom neural networks', *IEEE Access*, vol. 8, pp. 159790–159805. 5

Wang, W., Shen, J., Dong, X., Borji, A. & Yang, R., 2019, 'Inferring salient objects from human fixations.', *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, pp. 913–1927. 81

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M. & Summers, R. M., 2017b, 'ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases', *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3462–3471. xii, xiii, xiv, 2, 5, 6, 11, 19, 21, 22, 23, 24, 25, 28, 34, 43, 46, 47, 48, 51, 63, 67, 68, 69, 72, 79, 81

Yan, C., Yao, J., Li, R., Xu, Z. & Huang, J., 2018, 'Weakly supervised deep learning for thoracic disease classification and localization on chest x-rays', *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, BCB '18, Association for Computing Machinery, New York, NY, USA, p. 103–110, <`https://doi.org/10.1145/3233547.3233573`>. 5

Yao, L., Poblenz, E. & et. al., 2017, 'Learning to diagnose from scratch by exploiting dependency among labels', *arXiv preprint arXiv:1710.10501*. 2, 8, 11, 21, 28, 29, 34, 46, 55

Yao, L., Prosky, J., Poblenz, E., Covington, B. & Lyman, K., 2018, 'Weakly supervised medical diagnosis and localization from multiple resolutions', *arXiv preprint arXiv:1803.07703*. 7, 46, 48, 62, 69

Yue, K., Sun, M., Yuan, Y., Zhou, F., Ding, E. & Xu, F., 2018, 'Compact generalized non-local network', *Advances in Neural Information Processing Systems*, pp. 6511–6520. 62

Zhao, G., Qi, B. & Li, J., 2020, ‘Cross chest graph for disease diagnosis with structural relational reasoning.’, *arXiv preprint arXiv:2101.08992*. 5