Pattern Recognition Letters xxx (xxxx) xxx

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Multi-label chest X-ray image classification via category-wise residual attention learning

Qingji Guan [a,b], Yaping Huang [a,*]

[a] Beijing Key Laboratory of Traffic Data Analysis and Mining, Beijing Jiaotong University, No. 3 Shangyuancun, Beijing 100044, China
[b] Centre for Artificial Intelligence, University of Technology Sydney, 15 Broadway, Sydney, New South Wales 2007, Australia

ARTICLE INFO

Article history:
Available online xxx

Keywords:
Chest X-ray
Residual attention
Convolutional neural network
Image classification

ABSTRACT

This paper considers the problem of multi-label thorax disease classification on chest X-ray images. Identifying one or more pathologies from a chest X-ray image is often hindered by the pathologies unrelated to the targets. In this paper, we address the above problem by proposing a category-wise residual attention learning (*CRAL*) framework. *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. Specifically, the proposed framework consists of two modules: feature embedding module and attention learning module. The feature embedding module learns high-level features with a convolutional neural network (CNN) while the attention learning module focuses on exploring the assignment scheme of different categories. The attention module can be flexibly integrated into any feature embedding networks with end-to-end training. The comprehensive experiments are conducted on the Chest X-ray14 dataset. *CRAL* yields the average AUC score of 0.816 which is a new state of the art.

1. Introduction

Large-scale multimedia data (*i.e.*, image or video) has accelerated the development of tasks in computer vision, such as image retrieval and classification [1,6,18,19,24,40–43], video analysis [4,5,7,8,23,29]. In radiological studies, chest X-ray (CXR) is one of commonly used screening techniques in thorax disease diagnosis, such as nodules, cardiomegaly, effusion. Thousands of CXR images are captured in hospitals, making the computer-aided diagnosis (CAD) very important but challenging. Therefore, automatic analysis of CXR images would effectively assist clinical diagnosis and pathology finding. However, chest X-ray image analysis is a challenging task which suffers from the intrinsically complex relations of different pathologies. In this paper, we propose a category-wise residual attention learning framework that considers both the relevant and irrelevant categories for CXR image classification.

Commonly, CXR images are labeled with one or more pathologies, which makes the CXR image classification a multi-label problem. In the Chest X-ray14 dataset [37], 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. 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.

Deep learning has made noticeable progress in field of medical image analysis, such as classification [2,17,44], lesion segmentation or detection [10,22,27,47], image registration [3,21]. In this paper, 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. Fig. 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

* Corresponding author.
E-mail address: yphuang@bjtu.edu.cn (Y. Huang).

https://doi.org/10.1016/j.patrec.2018.10.027
0167-8655/

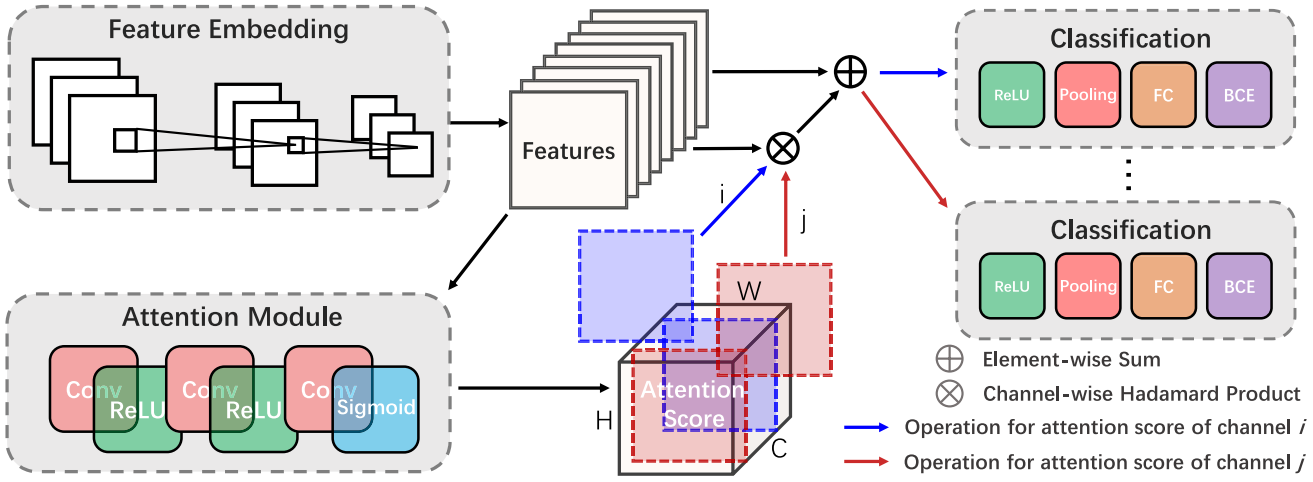

Fig. 1. Overview of the framework. There are two different attention mechanisms investigated in Section 3. Here, we take the first one *att1* as an example to illustrate the proposed framework. *CRAL* consists of two main modules. The feature embedding module is a CNN network which can be replaced by 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.

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. 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 Chest X-ray14 dataset. Experimental results demonstrate that our framework yields superior performance over the state-of-the-art approaches.

The rest of the paper is organized as follows. Section 2 introduces the related work in chest X-ray image classification and attention mechanism in medical image analysis. The proposed method is described in Section 3. We give the experimental results in Section 4. Section 5 concludes the whole work.

2. Related work

Image classification on the Chest X-ray14 dataset. Wang et al. [37] evaluate the classic CNN architectures, *i.e.*, AlexNet [16], VGGNet [31], GoogLeNet [33], ResNet [13], to predict the presence of multiple diseases. With a weakly manner proposed in [49], the disease lesion areas are located. Rajpurkar et al. [28] propose to classify the CXR image by fine-tuning a modified DenseNet [15] which replaces the last fully connected layer with a 14-output fully connected layer. Considering the relatively small size of lesion area in the whole image, Guan et al. [11] utilize an attention guided mask inference method to locate the region of interests (RoI) from the CNN feature maps. Combining with the global features, the RoIs are retrained to classify the image. Wang et al. [38] improve the classification performance with the aid of additional radiology reports. Both image and its corresponding clinical reports are embedded respectively and feed into a CNN-RNN model for training. Multi-level attentions are introduced to produce saliency-encoded text and image embeddings. Guendel et al. [12] propose to recognize the abnormality in the CXR image by a location-aware dense network. The spatial information in the high-resolution images is utilized during training and testing. Tang et al. [34] identify the disease category and localize the lesion areas through an attention-guided curriculum learning method. Severity-level attributes mined from radiology reports are leveraged. Shen et al. [30] accomplish the disease identification and localization by combining the routing-by agreement mechanism and the deep convolutional neural network.

From the view of standard multi-label image classification, researchers attempt to capture the relationships among image labels. Kumar et al. [17] 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. [44] take a variant of DenseNet as image encoder and capture the label correlation among the 14 pathologies with a Long-short Term Memory Network (LSTM) [14] which performs as a decoder. This paper 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.

Some researchers also focus on locating the position of lesion area with image-level or limited supervision. Li et al. [20] propose to jointly learn the disease identification and localization with the supervision of image-level label and limited lesion area location information. Through combining the loss on the image with/without bounding box, end-to-end training is performed. With bounding box supervision, the classification performance is further improved. Yao et al. [45] consider multiple feature resolutions with both residual and dense connections to generate image saliency maps. And a learnable lower-bounded adaption method is used to build a sharpness prior and localize abnormal areas with image-level labels.

Attention models in chest X-ray image analysis. Attention models have been successfully explored in field of natural language processing [35,39], computer vision [9,32,48] and so on. The

Please cite this article as: Q. Guan, Y. Huang, Multi-label chest X-ray image classification via category-wise residual attention learning, Pattern Recognition Letters (2018), https://doi.org/10.1016/j.patrec.2018.10.027

multi-label CXR image classification problem needs to learn the discriminative features to distinguish different pathologies. Commonly, one pathology is often semantically annotated to the lesion area, which is the critical cues for classification and localization. Ypsilantis et al. [46] propose a stochastic attention-based model to determine which regions should be visually explored, and conclude whether a specific radiological abnormality exists or not. However, only one disease "*enlarged heart*" is considered in their work. Recently, Pesce et al. [26] explore a soft attention mechanism from the saliency map of CNN features to locate lung nodule in radiography, and a localization loss is calculated by comparing the predicted position with the annotated position. In this paper, *CRAL* integrates the residual attention in [36] and multi-label attention in [50] into a category-wise residual attention module to classify the chest X-ray images. No any other annotation information is required.

3. 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 3.1. Then, the feature embedding module and the residual attention module are introduced in Section 3.2 and Section 3.3, respectively. We finally present the optimizing strategy in Section 3.4.

3.1. Architecture of CRAL

The architecture of *CRAL* is presented in Fig. 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 [13,36], 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 $\mathbf{L} = [l_1, l_2, \ldots, l_C]$ 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 $\mathbf{L}$ is a zero vector, it means that none of all pathologies exists in the image.

3.2. 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 [13] or DenseNet-121 [15] network as the backbone. Next, we take ResNet-50 as an example to introduce the feature embedding module.

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,

$$F = f_{cnn}(I; \theta_{cnn}), F \in \mathbb{R}^{H \times W \times N} \quad (1)$$

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.

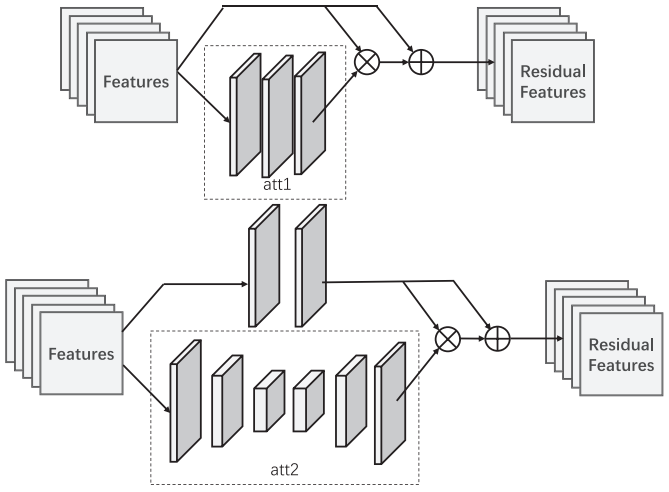

Fig. 2. Architecture of residual attention module (with *att1* and *att2*). *att1* consists of two 3×3 convolutional layers followed by ReLU, one 1×1 convolutional layer and one non-linear activation layer (Sigmoid). For *att2*, the input CNN features F are fed into the "hourglass" attention branch and a convolutional branch, respectively. Through the channel-wise Hadamard product and element-wise sum operations, a residual feature is formed by the learned features $\tilde{\mathbf{F}}$ and its weighted version $\mathbf{A} \odot \tilde{\mathbf{F}}$.

3.3. 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 the residual attention module which are denoted as *att*1 and *att*2, respectively. The architectures of residual attention are presented in Fig. 2. *att1* consists of two 3×3 convolutional layers and each followed by a 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 [36]. 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 upsampling 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 Fig. 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 *att*1 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

Please cite this article as: Q. Guan, Y. Huang, Multi-label chest X-ray image classification via category-wise residual attention learning, Pattern Recognition Letters (2018), https://doi.org/10.1016/j.patrec.2018.10.027

aim to automatically predict label attention scores for each class,

$$Z = f_{att}(F; \theta_{att}), Z \in \mathbb{R}^{H\times W\times C} \tag{2}$$

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,

$$a^c_{i,j} = \frac{1}{1+exp(-z^c_{i,j})}, A \in \mathbb{R}^{H\times W\times C} \tag{3}$$

where $a^c_{i,j}$ and $z^c_{i,j}$ 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 *att*1 as an example to illustrate the remain of the attention module in the following section. The category-wise weighted CNN features are denoted as $V = \{V^1, V^2, \ldots, V^C\}$, where $V^c = \{v^{1,c}, v^{2,c}, \ldots, v^{N,c}\}$. 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,

$$v^{n,c}_{i,j} = F^n_{i,j} \odot a^c_{i,j}, v^{n,c} \in \mathbb{R}^{H\times W} \tag{4}$$

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

$$H^{n,c}_{i,j} = F^n_{i,j} + V^{n,c}_{i,j} = (\mathbf{1} + a^c_{i,j}) \odot F^n_{i,j}, \tag{5}$$

where $a^c_{i,j}$ ranges in [0,1], and it works as feature selectors which enhance discriminative features and suppress irrelevant features. $\mathbf{1}$ 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,

$$\tilde{H}^c = max(0, H^c) \tag{6}$$

and

$$\bar{H}^{n,c} = \frac{1}{K}\sum_{i,j}\tilde{H}^{n,c}_{i,j}, \tag{7}$$

where K is the number of activation values in $\tilde{H}^{n,c}$. $\bar{H}^c = \{\bar{H}^{1,c}, \bar{H}^{2,c}, \ldots, \bar{H}^{N,c}\}$ is a N-dim vector. For *att*2, due to the CNN features from the feature embedding module are fed into two residual blocks, the F in Eq. 4 and Eq. 5 should be replaced by the new features $\tilde{F}$. We discuss these two attention mechanisms in Section 4.

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.

3.4. 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,

$$\hat{H}^c = f_{cls}(\bar{H}^c; \theta^c), \tag{8}$$

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

$$\tilde{p}(c|\hat{H}^c) = \frac{1}{1+exp(-p(c|\hat{H}^c))}, \tag{9}$$

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

$$\mathcal{L}(W) = -\frac{1}{C}\sum_{c=1}^{C} l_c \log(\tilde{p}(c|\hat{H}^c)) + (1-l_c)\log(1-\tilde{p}(c|\hat{H}^c)), \tag{10}$$

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

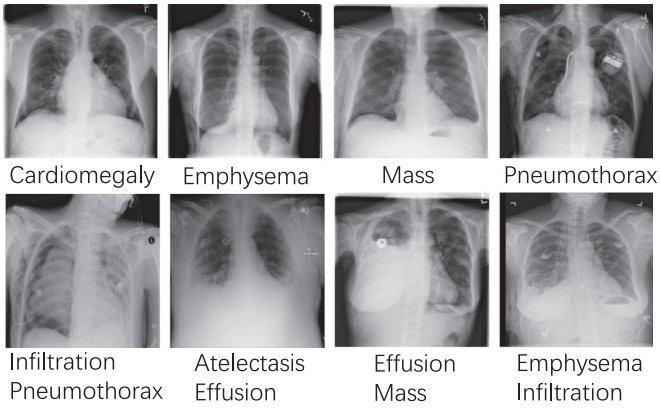

Fig. 3. Example images and the corresponding labels in the Chest X-ray14 dataset. Each image is labeled with one or more pathologies.

4. 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 discusses different attention mechanisms and demonstrates the performance of the proposed *CRAL* framework. The ablation study is presented to show the efficiency of *CRAL* in Section 4.4. At last, we visualize some feature heatmaps with *CRAL* and some classification results in Section 4.5.

4.1. Dataset and evaluation metric

Dataset. We evaluate the *CRAL* framework on a large scale chest X-ray dataset, Chest X-ray14[1], released by NIH [37]. 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"). Fig. 3 shows some examples and the corresponding annotations in Chest X-ray14.

Evaluation metric. In our experiment, we utilize the dataset split provided by [37]. There is no any patients overlap in train and test subsets. Each image is labeled with a one-shot vector

[1] https://nihcc.app.box.com/v/ChestXray-NIHCC

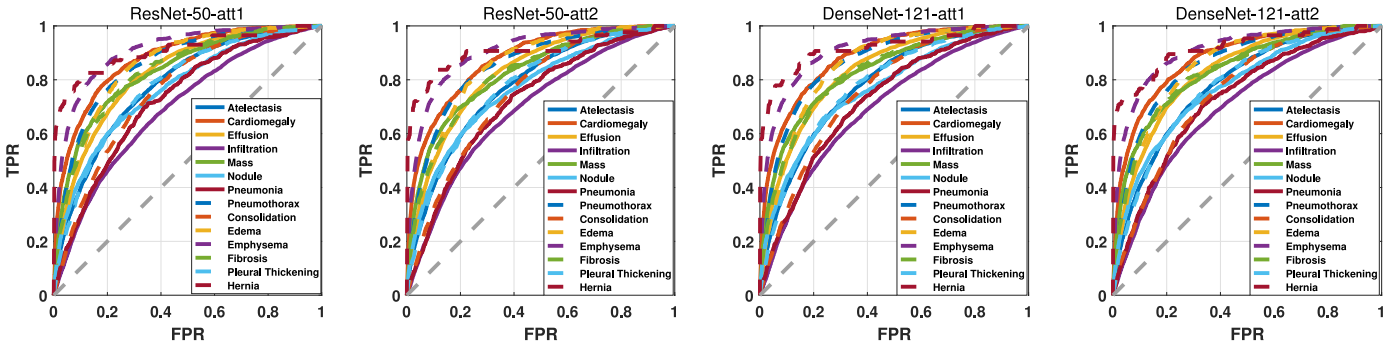

Fig. 4. ROC curves of four combinations of CNN backbones and attention mechanisms (ResNet-50-att1, ResNet-50-att2, DenseNet-121-att1, and DenseNet-121-att2) over the 14 pathologies. The corresponding AUC scores are given in Table 1.

Table 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. ResNet-50 (R-50) and Desnet-121 (D-121) are used as backbones in our approach. For each column, the best results are highlighted in bold. * represents that the combination of ResNet and DenseNet is used in Yao et al. [45]. – represents the network used in the corresponding reference is not illustrated.

Method	CNN	Atel	Card	Effu	Infi	Mass	Nodu	Pne1	Pne2	Cons	Edem	Emph	Fibr	PT	Hern	Mean
Wang et al. [37]	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. [12]	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. [45]	*	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. [20]	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. [20]	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 et al. [30]	–	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. [34]	–	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.

$\mathbf{L} = [l_1, l_2, \ldots, l_C]$, C is 14 in Chest X-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.

4.2. 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 [25].

4.3. Evaluation

We evaluate our method on the Chest X-ray14 dataset. ResNet-50 [13] and DenseNet-121 [15] 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 *att*1 and *att*2 to validate the effectiveness of the proposed *CRAL* framework. The results are summarized in Table 1, Table 2 and Fig. 4.

Table 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

Both *att1* and *att2* improve the classification performance on the chest X-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 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 1 and Fig. 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 [44], [17] or [28], 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 [37] with image-level supervision. We evaluate *CRAL* and

Please cite this article as: Q. Guan, Y. Huang, Multi-label chest X-ray image classification via category-wise residual attention learning, Pattern Recognition Letters (2018), https://doi.org/10.1016/j.patrec.2018.10.027

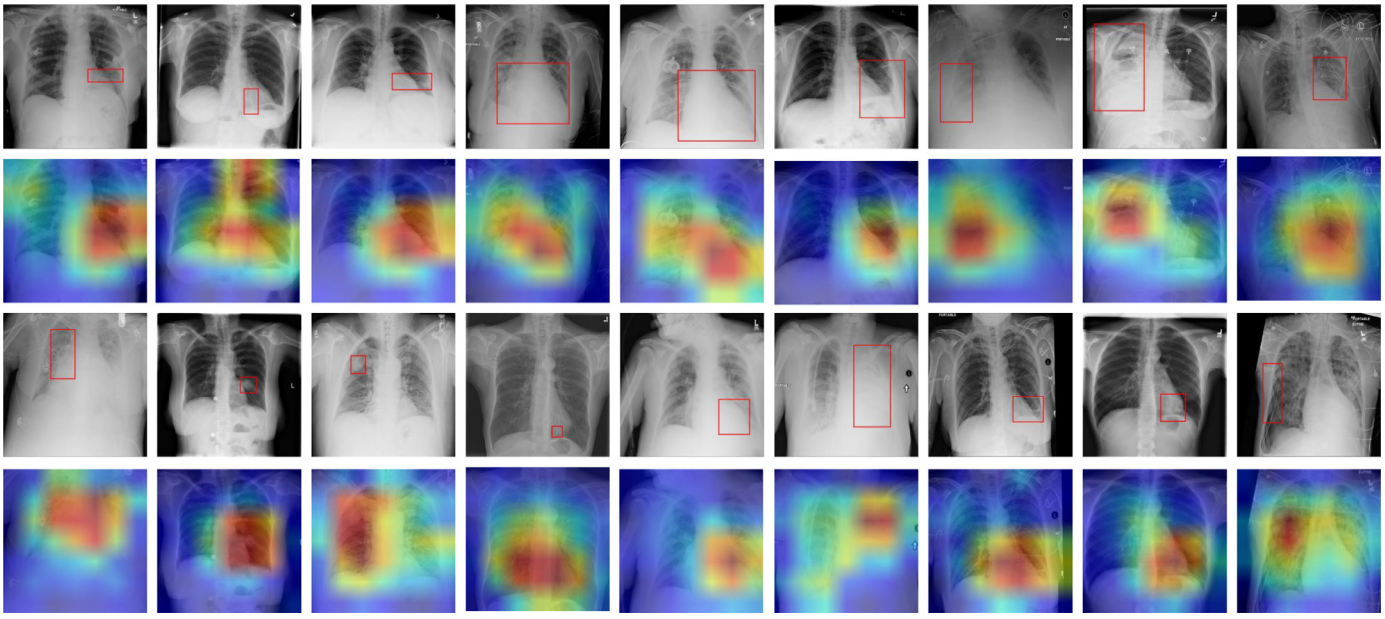

Fig. 5. Examples of heatmaps generated from the learned features (from ResNet-50). The ground truth bounding boxes provided by [37] are annotated on the original images. Note that the heatmaps are zoomed to the same size as the input images, and the heatmaps may be a few difference due to the usage of random cropping in testing.

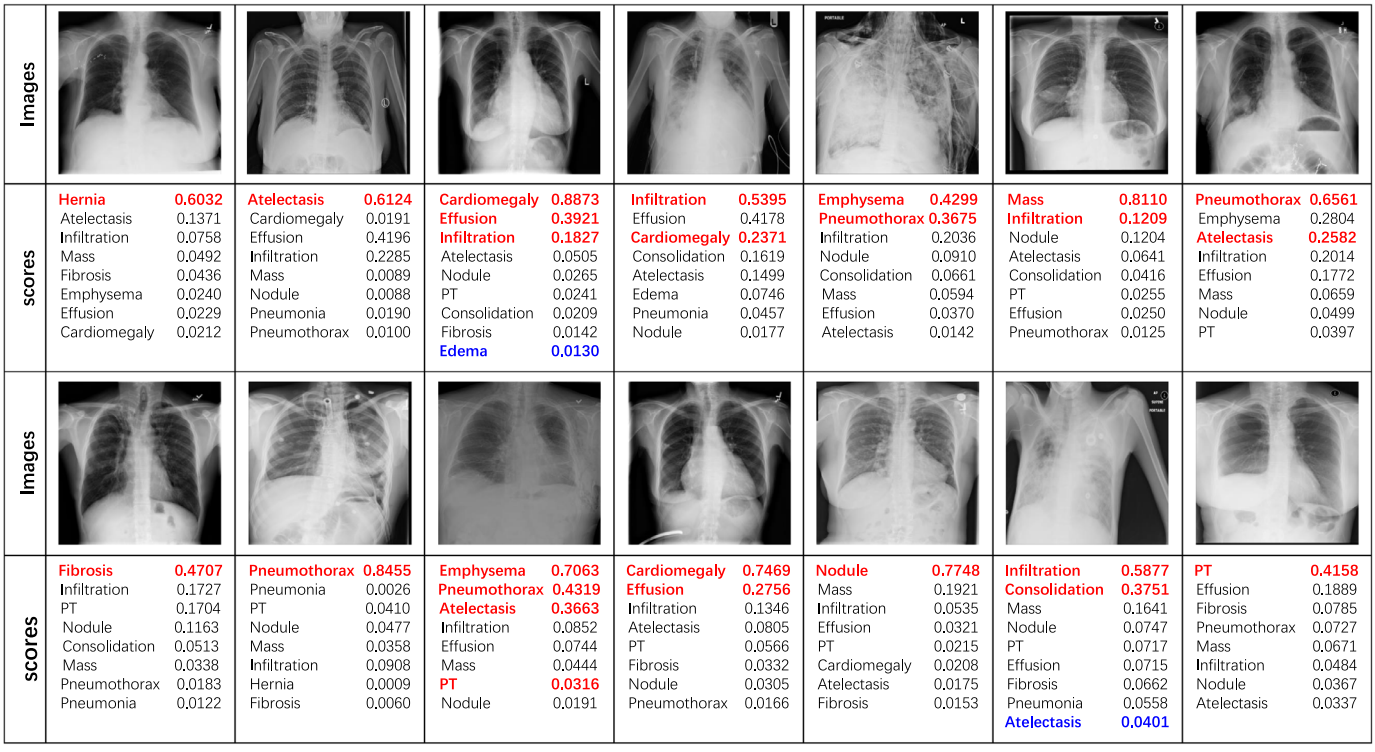

Fig. 6. Examples of classification results. We present the top-8 predicted categories and the corresponding probability scores. The ground truth labels are highlighted in red or blue.

compare it with state-of-the-art methods on the Chest X-ray14 dataset. The results are summarized in Table 1 and Fig. 4. Wang et al. [37] 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. [12] 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 [12] 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

Please cite this article as: Q. Guan, Y. Huang, Multi-label chest X-ray image classification via category-wise residual attention learning, Pattern Recognition Letters (2018), https://doi.org/10.1016/j.patrec.2018.10.027

other extra data. Yao et al. [45] achieve the chest X-ray image classification and localization with a multiple resolutions setting. Li et al. [20] utilize additional lesion area annotation as supervision. Tang et al. [34] progressively learn an attention-guided curriculum to identify the pathologies and the attributes mining from radiology reports are used. Shen et al. [30] 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 [37] and [20] with 7.1% and 7.7%. With DenseNet-121, it surpasses the previous state of the art [12] 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 [12] about 2% (DenseNet-121 with *att2*). The ROC curves of 14 pathologies with ResNet-50 and DenseNet-121 are presented in Fig. 4. In all, the classification performance reported in this paper compares favorably against previous methods.

4.4. Ablation study

To evaluate the effectiveness of the components of residual attention module, we conduct additional ablation experiments on the Chest X-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 2. The performance of *CRAL* over 14 pathologies is reported in Table 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.

4.5. Qualitative results

We visualize some feature heatmaps and classification results shown in Fig. 5 and Fig. 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 ResNet-50), and then count the maximum values along feature channels. In Fig. 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 Fig. 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.

5. Conclusion

In this paper, we propose 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. In the future, due to the expensive cost of annotating the position of lesion areas, accurate localization method with weak supervision will be investigated.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 61273364, 61473031 & 61472029) and the Fundamental Research Funds for the Central Universities (2016JBZ005, 2016YJS041, 2018YJS045, 2018YJS035). We thank Qi Zou and Mengyang Pu for their valuable suggestions on our manuscript.

References

[1] Z. Ahmadi, S. Kramer, A label compression method for online multi-label classification, Pattern Recognit. Lett. (2018).
[2] M. Anthimopoulos, S. Christodoulidis, L. Ebner, A. Christe, S. Mougiakakou, Lung pattern classification for interstitial lung diseases using a deep convolutional neural network, IEEE Trans. Med. Imaging 35 (2016) 1207–1216.
[3] H. Bülow, L. Dooley, D. Wermser, Application of principal axes for registration of NMRimage sequences, Pattern Recognit. Lett. 21 (2000) 329–336.
[4] X. Chang, Z. Ma, M. Lin, Y. Yang, A.G. Hauptmann, Feature interaction augmented sparse learning for fast kinect motion detection, IEEE Trans. Image Process. 26 (2017) 3911–3920.
[5] X. Chang, Z. Ma, Y. Yang, Z. Zeng, A.G. Hauptmann, Bi-level semantic representation analysis for multimedia event detection, IEEE Trans. Cybern. 47 (2017) 1180–1197.
[6] X. Chang, Y. Yang, Semisupervised feature analysis by mining correlations among multiple tasks, IEEE Trans. Neural. Netw. Learn Syst. 28 (2017) 2294–2305.
[7] X. Chang, Y. Yu, Y. Yang, E.P. Xing, Semantic pooling for complex event analysis in untrimmed videos, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2017) 1617–1632.
[8] R. De Rosa, I. Gori, F. Cuzzolin, N. Cesa-Bianchi, Active incremental recognition of human activities in a streaming context, Pattern Recognit. Lett. 99 (2017) 48–56.
[9] J. Fu, H. Zheng, T. Mei, Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4476–4484.
[10] G. Gilanie, M. Attique, S. Naweed, E. Ahmed, M. Ikram, et al., Object extraction from t2 weighted brain mr image using histogram based gradient calculation, Pattern Recognit. Lett. 34 (2013) 1356–1363.
[11] Q. Guan, Y. Huang, Z. Zhong, Z. Zheng, L. Zheng, Y. Yang, Diagnose like a radiologist: attention guided convolutional neural network for thorax disease classification, arXiv preprint arXiv:1801.09927 (2018).
[12] S. Guendel, S. Grbic, B. Georgescu, K. Zhou, L. Ritschl, A. Meier, D. Comaniciu, Learning to recognize abnormalities in chest x-rays with location-aware dense networks, arXiv preprint arXiv:1803.04565 (2018).
[13] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[14] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (1997) 1735–1780.
[15] G. Huang, Z. Liu, K.Q. Weinberger, L. van der Maaten, Densely connected convolutional networks, Proceed. IEEE Conf. Comp.Vision Pattern Recog. (2017) 4700–4708.
[16] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.

[17] P. Kumar, M. Grewal, M.M. Srivastava, Boosted cascaded convnets for multilabel classification of thoracic diseases in chest radiographs, arXiv preprint arXiv:1711.08760 (2017).
[18] F. Li, M. Zhang, B. Tian, B. Chen, G. Fu, D. Ji, Recognizing irregular entities in biomedical text via deep neural networks, Pattern Recognit Lett (2017).
[19] Z. Li, F. Nie, X. Chang, Y. Yang, Beyond trace ratio: weighted harmonic mean of trace ratios for multiclass discriminant analysis, IEEE Trans. Knowl. Data Eng. 29 (2017) 2100–2110.
[20] Z. Li, C. Wang, M. Han, Y. Xue, W. Wei, L.-J. Li, F.-F. Li, Thoracic disease identification and localization with limited supervision, in: CVPR, 2018, pp. 8290–8299.
[21] R. Liao, S. Miao, P. de Tournemire, S. Grbic, A. Kamen, T. Mansi, D. Comaniciu, An artificial agent for robust image registration., in: AAAI, 2017, pp. 4168–4175.
[22] P. Liskowski, K. Krawiec, Segmenting retinal blood vessels with deep neural networks, IEEE Trans. Med. Imag. 35 (2016) 2369–2380.
[23] D.C. Luvizon, H. Tabia, D. Picard, Learning features combination for human action recognition from skeleton sequences, Pattern Recognit. Lett. 99 (2017) 13–20.
[24] Z. Ma, X. Chang, Y. Yang, N. Sebe, A.G. Hauptmann, The many shades of negativity, IEEE Trans. Multimedia 19 (2017) 1558–1568.
[25] A. Paszke, S. Gross, S. Chintala, G. Chanan, Pytorch, 2017.
[26] E. Pesce, P.-P. Ypsilantis, S. Withey, R. Bakewell, V. Goh, G. Montana, Learning to detect chest radiographs containing lung nodules using visual attention networks, arXiv preprint arXiv:1712.00996 (2017).
[27] V. Rajinikanth, S.C. Satapathy, S.L. Fernandes, S. Nachiappan, Entropy based segmentation of tumor from brain mr images–a study with teaching learning based optimization, Pattern Recognit. Lett. 94 (2017) 87–95.
[28] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, et al., Chexnet: radiologist-level pneumonia detection on chest x-rays with deep learning, arXiv preprint arXiv:1711.05225 (2017).
[29] M. Ribeiro, A.E. Lazzaretti, H.S. Lopes, A study of deep convolutional auto-encoders for anomaly detection in videos, Pattern Recognit. Lett. (2017).
[30] Y. Shen, M. Gao, Dynamic routing on deep neural network for thoracic disease classification and sensitive area localization, in: International Workshop on Machine Learning in Medical Imaging, Springer, 2018, pp. 389–397.
[31] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).
[32] Y. Sun, R. Fisher, Object-based visual attention for computer vision, Artif. Intell. 146 (2003) 77–123.
[33] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
[34] Y. Tang, X. Wang, A.P. Harrison, L. Lu, J. Xiao, R.M. Summers, Attention-guided curriculum learning for weakly supervised classification and localization of thoracic diseases on chest radiographs, in: International Workshop on Machine Learning in Medical Imaging, Springer, 2018, pp. 249–258.
[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: NIPS, 2017, pp. 6000–6010.
[36] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, X. Tang, Residual attention network for image classification, in: CVPR, 2017, pp. 3156–3164.
[37] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R.M. Summers, Chest X-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3462–3471.
[38] X. Wang, Y. Peng, L. Lu, Z. Lu, R.M. Summers, Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays, in: CVPR, 2018, pp. 9049–9058.
[39] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: ICML, 2015, pp. 2048–2057.
[40] Y. Yan, F. Nie, W. Li, C. Gao, Y. Yang, D. Xu, Image classification by cross-media active learning with privileged information, IEEE Trans. Multimedia 18 (2016) 2494–2502.
[41] Y. Yang, Z. Ma, A.G. Hauptmann, N. Sebe, Feature selection for multimedia analysis by sharing information among multiple tasks, IEEE Trans. Multimedia 15 (2013) 661–669.
[42] Y. Yang, Z. Ma, F. Nie, X. Chang, A.G. Hauptmann, Multi-class active learning by uncertainty sampling with diversity maximization, Int. J. Comput. Vis. 113 (2015) 113–127.
[43] Y. Yang, F. Nie, D. Xu, J. Luo, Y. Zhuang, Y. Pan, A multimedia retrieval framework based on semi-supervised ranking and relevance feedback, IEEE Trans. Pattern Anal. Mach. Intell. 34 (2012) 723–742.
[44] L. Yao, E. Poblenz, D. Dagunts, B. Covington, D. Bernard, K. Lyman, Learning to diagnose from scratch by exploiting dependencies among labels, arXiv preprint arXiv:1710.10501 (2017).
[45] L. Yao, J. Prosky, E. Poblenz, B. Covington, K. Lyman, Weakly supervised medical diagnosis and localization from multiple resolutions, arXiv preprint arXiv:1803.07703 (2018).
[46] P.-P. Ypsilantis, G. Montana, Learning what to look in chest x-rays with a recurrent visual attention model, arXiv preprint arXiv:1701.06452 (2017).
[47] Y. Yuan, M. Chao, Y.-C. Lo, Automatic skin lesion segmentation using deep fully convolutional networks with jaccard distance, IEEE Trans. Med. Imaging 36 (2017) 1876–1886.
[48] H. Zheng, J. Fu, T. Mei, J. Luo, Learning multi-attention convolutional neural network for fine-grained image recognition, in: ICCV, pp. 5209–5217.
[49] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2921–2929.
[50] F. Zhu, H. Li, W. Ouyang, N. Yu, X. Wang, Learning spatial regularization with image-level supervisions for multi-label image classification, in: CVPR, 2017, pp. 5513–5522.

Please cite this article as: Q. Guan, Y. Huang, Multi-label chest X-ray image classification via category-wise residual attention learning, Pattern Recognition Letters (2018), https://doi.org/10.1016/j.patrec.2018.10.027