Pattern Recognition Letters 131 (2020) 38–45

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Thorax disease classification with attention guided convolutional neural network

Qingji Guan[a,b], Yaping Huang[a,*], Zhun Zhong[c,b], Zhedong Zheng[b], Liang Zheng[b], Yi Yang[b]

[a] Beijing Key Laboratory of Traffic Data Analysis and Mining, Beijing Jiaotong University, No. 3 Shangyuancun, Beijing 100044, China
[b] ReLER Lab, Centre for Artificial Intelligence, University of Technology Sydney, 15 Broadway, Sydnedy, NSW 2007, Australia
[c] Department of Artifical Intelligence, Xiamen University, China

ARTICLE INFO

Article history:
Received 5 May 2019
Revised 26 October 2019
Accepted 29 November 2019
Available online 30 November 2019

Keywords:
CXR image classification
Visual attention
Feature ensemble

ABSTRACT

This paper considers the task of thorax disease diagnosis on chest X-ray (CXR) images. Most existing methods generally learn a network with global images as input. However, thorax diseases usually happen 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 the network performance. For addressing the above problems, we propose to integrate the global and local cues into a three-branch attention guided convolution neural network (AG-CNN) to identify thorax diseases. An attention guided mask inference based cropping strategy is proposed to avoid noise and improve alignment in the global branch. AG-CNN also integrates the global cues to compensate the lost discriminative cues by the local branch. Specifically, we first learn a global CNN branch using global images. Then, guided by the attention heatmap generated from the global branch, we infer 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. 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. Introduction

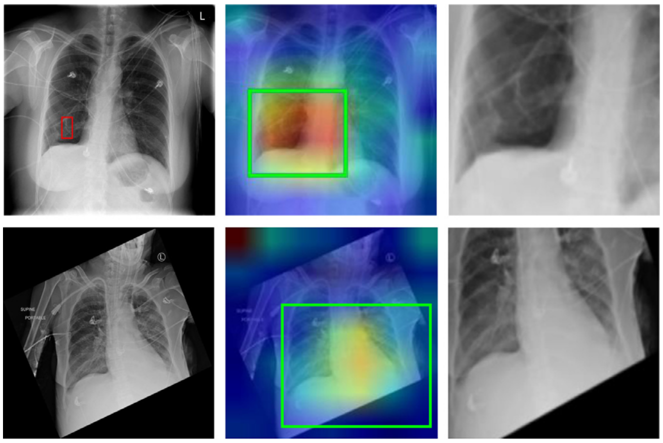
The chest X-ray (CXR) has been one of the most common radiological examinations in lung and heart disease diagnosis. Currently, reading CXRs mainly relies on professional knowledge and careful manual observation. Due to the complex pathologies and subtle texture changes of different lung lesion in images, radiologists may make mistakes even when they have experienced long-term clinical training and professional guidance. Therefore, it is of importance to develop the CXR image classification methods to support clinical practitioners. The noticeable progress in deep learning has benefited many trials in medical image analysis. In this paper, we investigate the CXR classification task using deep learning.

Several existing works on CXR classification typically employ the *global image* for training. For example, Wang *et al.* [25] evaluate four classic CNN architectures [6,10,22,23] to tell the presence of multiple pathologies using a global CXR image. Viewing CXR classification as a multi-label recognition problem, Yao et al. [29] explore the correlation among the 14 pathologic labels with global images in ChestX-ray14 [25]. Using a variant of DenseNet [8] as an image encoder, they adopt the Long-short Term Memory Networks (LSTM) [7] to capture the dependencies. Kumar et al. [12] investigate that which loss function is more suitable for training CNNs from scratch and present a boosted cascaded CNN for global image classification. The recent effective method consists in CheXNet [19], which fine-tunes a modified 121-layer DenseNet on the global chest X-ray images.

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. 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 [2], where the object of interest is usually positioned in the image center. 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

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

https://doi.org/10.1016/j.patrec.2019.11.040
0167-8655/© 2019 Elsevier B.V. All rights reserved.



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

Fig. 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).

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. 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 paper, 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 paper 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.

Our 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 achieves competitive accuracy.

2. Related works

The problem of Chest X-ray image classification has been extensively explored in the field of medical image analysis. Recently, Wang *et al.* [25] release the ChestX-ray14 dataset, which is the largest chest X-ray dataset by far. ChestX-ray14 collects 112,120 frontal-view chest X-ray images of 30,805 unique patients. Each radiography is labeled with one or more of 14 common thorax diseases. It is also large enough for deep learning, so we adopt it for performance evaluation.

Deep learning for chest X-ray image analysis. Deep networks have been explored and succeeded in various tasks of computer vision [27,30,31]. Recent surveys [3,14,18] have demonstrated that deep learning technologies have been extensively applied to the field of medical image analysis [11,20], especially in chest X-ray image classification [4,19,25]. Yao et al. [29] and Kumar et al. [12] classify the chest X-ray images by investigating the potential dependencies among the labels from the aspect of multi-label problems. Rajpurkar *et al.* [19] train a convolutional neural network to address the multi-label classification problem. With the aid of additional radiology reports, Wang *et al.* [26] improve the chest X-ray image classification performance with saliency-encoded text and image embeddings. Guendel *et al.* [5] propose a location-aware dense network to recognize the abnormality in the CXR image. This paper departs from the previous methods in that we make use of the attention mechanism and fuse the local and global information to improve the classification performance without auxiliary medical report or lesion position.

Global-local strategy in other domains. Combining the global and local cues has been explored in tasks of different domains, such as document analysis [1], object detection and recognition [15,17], image retrieval [9,21], and natural image classification[16,28,32]. Akbari *et al.* [1] propose an adaptive multi-modal multi-view ranking model to jointly regularize the relatedness among modalities, the effects of feature views extracted from different modalities, as well as the complex relations among multi-modal documents. Luo *et al.* [15] propose to detect the salient object with a simplified convolutional neural network which combines local and global information through a multi-resolution 4 × 5 grid structure. Shyu *et al.* [21] try to access the utility of localized versus global features for the domain of HRCT images of the lung. Yang *et al.* [28] design a network for learning the localized informative regions in a self-supervision mechanism for fine-grained recognition. However, in the field of chest X-ray image analysis, the global-local strategy has not been well explored. In this paper, we propose a chest X-ray recognition approach in analogy to the human expert in terms of attention. We

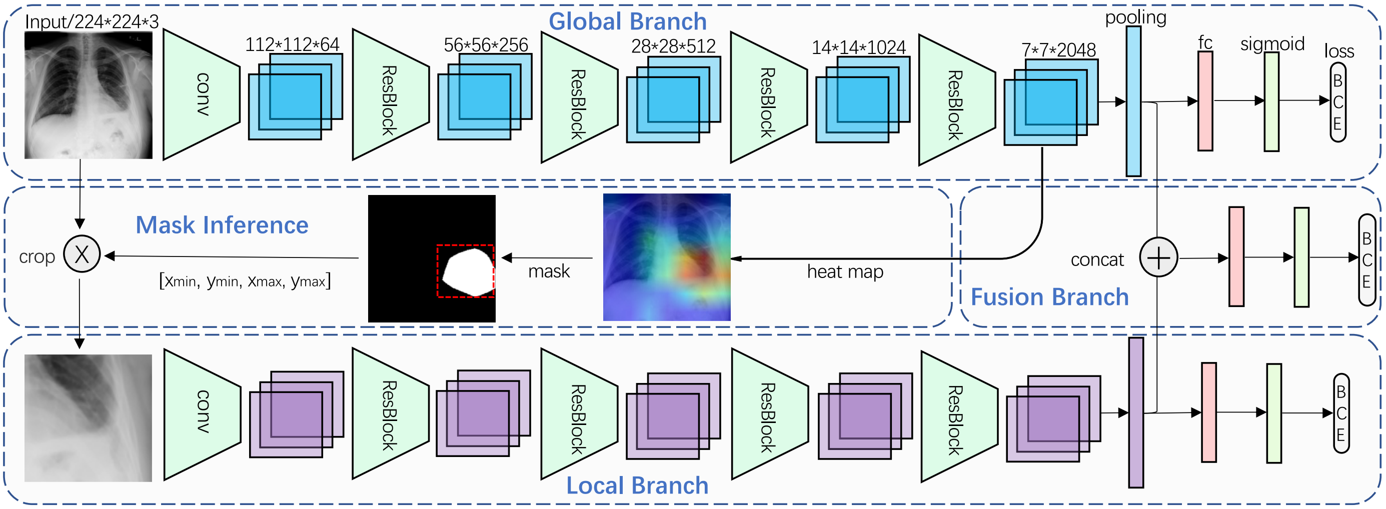


Fig. 2. Overall framework of the attention guided convolutional neural network (AG-CNN, showing ResNet-50 as backbone). "BCE" represents binary cross entropy loss. The spatial resolution of heatmap generated from the last convolutional layer of the global branch is 7 × 7. Then we resize the heatmap to 224 × 224 by bilinear interpolation. The input image is added to the heatmap for visualization.

focus on exploring both the global and the informative local cues to reach a final decision of chest X-ray images. In this sense, the proposed method contributes to mimicking the human diagnosing procedure and reporting competitive accuracy.

Attention models in medical image analysis. The CXR classification problem needs to tell the relatively subtle differences between different diseases. Usually, a disease is often characterized by a lesion region, which contains critical cues for classification. Tang *et al.* [24] 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. Guan *et al.* [4] introduce a category-wise attention learning method which aims to strengthen the relevant features and suppress the irrelevant features for chest X-ray image classification. In this paper, AG-CNN locates the salient regions with an attention guided mask inference process, and learns the discriminative feature for classification. Compared with the method which relies on bounding box annotations, our method only needs image-level labels without any extra information.

3. The proposed approach

3.1. Structure of AG-CNN

The architecture of AG-CNN is presented in Fig. 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 information derived from the global image as input. In the global branch, we train a variant of ResNet-50 [6] 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

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

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

$$\mathcal{L}(W_g) = -\frac{1}{C}\sum_{c=1}^{C} l_c log(\widetilde{p}_g(c|I)) + (1-l_c) log(1-\widetilde{p}_g(c|I)), \tag{2}$$

where l_c is the groundtruth label of the cth 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. (2).

3.2. Attention guided mask inference

In this paper, we construct a binary mask 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 kth channel of the output of the last convolutional layer, where $k \in \{1, \ldots, K\}$, $K = 2048$ 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,

$$H_g(x, y) = \max_k(|f_g^k(x, y)|), k \in \{1, \ldots, K\}. \tag{3}$$

The values in H_g directly indicate the importance of the activations for classification. In Figs. 1(b) and 3 (the second row), some ex-

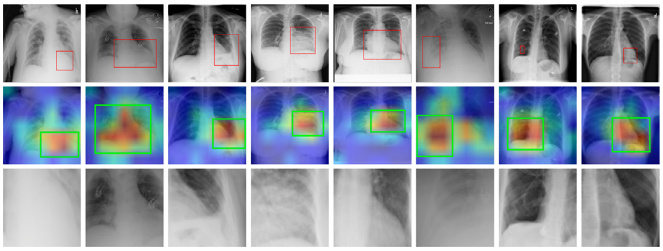


Fig. 3. The process of lesion area generation. (**Top:**) global CXR images of various thorax diseases for the global branch. Note that we do not use the bounding boxes for training or testing. (**Middle:**) corresponding visual examples of the output of the mask inference process. Higher/lower response is denoted with red/blue. (**Bottom:**) cropped and resized images from the green bounding boxes which are fed to the local branch. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

amples 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 result in subtle numerical differences in heatmap, but may not effect the classification significantly. Therefore, we compute the heatmap with Eq. (3) in our experiment. The comparison of these statistics is presented in Section 4.2.

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,

$$M(x,y) = \begin{cases} 1, & H_g(x,y) > \tau \\ 0, & \text{otherwise} \end{cases}, \tag{4}$$

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 $\tau = 0.7$ in Fig. 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).

3.3. Training strategy of AG-CNN

This paper adopts a three-stage training scheme 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. 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. (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. (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 4.2.

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. (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. (2) (Stage II);
Concentrating $Pool_g$ and $Pool_l$, learning $W_f$, computing $\widetilde{p}_f(c|[I, I_c])$, optimizing by Eq. (2).
```

4. Experiment

Dataset. We evaluate the AG-CNN framework using the ChestX-ray14 [25]. 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 $\mathbf{L} = [l_1, l_2, \ldots, l_c, \ldots, l_C]$ in which $l_c \in \{0, 1\}, C = 15$. l_{15} represents the label with “No Finding”.

4.1. 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 4.3, 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).

4.2. Evaluation

We evaluate our method on the ChestX-ray14 dataset. Mostly, ResNet-50 [6] is used as backbone, but the AUC and ROC curve obtained by DenseNet-121 [8] 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 1 and Fig. 7. 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, these are competitive 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 [19]. Moreover, we observe that Infiltration has the lower recognition accuracy (0.728 and 0.717 for ResNet-50 and DenseNet-121, respectively). This is because the diagnosis of Infiltration mainly relies on the texture change among the lung area, which is challenging to recognize. The pathology Cardiomegaly achieves higher recognition accuracy (0.904 and 0.912

Table 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.* [25]	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.* [29]	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.* [19]	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.* [12]	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.

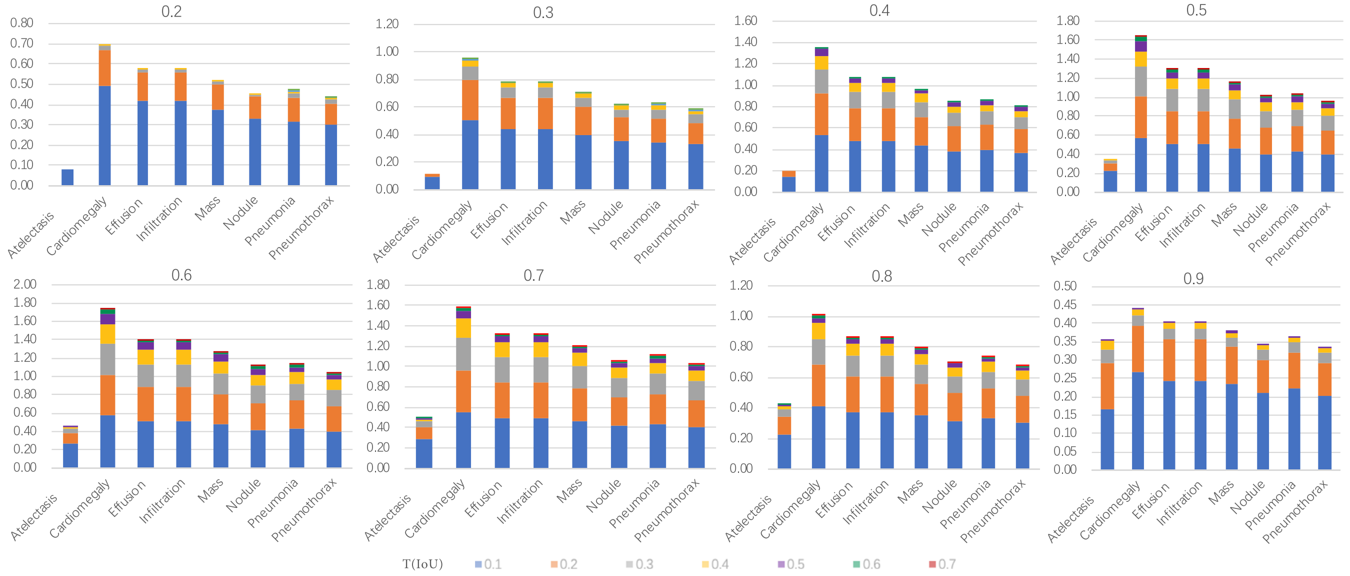


Fig. 4. The localization accuracy of different threshold of τ. Each sub-figure is the accuracy for different τ. And in each sub-figure, different color represents the threshold of IoU (T(IoU)) when measuring the accuracy of the predicted bounding box. Better view as zoomed.

for ResNet-50 and DenseNet-121, respectively), which is characterized by the relative solid region (heart).

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 1 and Fig. 7.

Using ResNet-50 and DenseNet-121, the average AUC score is 0.817 and 0.810, respectively, which is higher than [12,25]. 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 [25]. 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. 4 presents the localization accuracy of different τ in $\{0.2, 0.3, \ldots, 0.9\}$. In each sub-figure, different color represents the threshold of IoU when measuring the accuracy of the predicted bounding box. As shown in Fig. 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 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 2. Under different IoU thresholds, the localization accuracy of our method is consistently higher than [25]. Because both our method and [25] only use image-level labels, this comparison could be regarded as fair. Compared with [13], our method is inferior. The reason is that [13] uses additional ground truth bounding boxes which we do not use. Therefore, it is expected that [13] 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”

Table 2
Comparison of localization accuracy.

T(IoU)	Model	Atel	Card	Effu	Infi	Mass	Nodu	Pne1	Pne2	mean
0.1	[25]	0.69	0.94	0.66	0.71	0.40	0.14	**0.63**	0.38	0.57
	[13]	**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	[25]	0.47	0.68	0.45	0.48	0.26	0.05	0.35	0.23	0.37
	[13]	**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	[25]	0.24	0.46	0.30	0.28	0.15	0.04	0.17	0.13	0.22
	[13]	**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	[25]	0.09	0.28	0.20	0.12	0.07	0.01	0.08	0.07	0.12
	[13]	**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	[25]	0.05	0.18	0.11	0.07	0.01	0.01	0.03	0.03	0.06
	[13]	**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	[25]	0.02	0.08	0.05	0.02	0.00	0.01	0.02	0.03	0.03
	[13]	**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	[25]	0.01	0.03	0.02	0.00	0.00	0.00	0.01	0.02	0.01
	[13]	**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 [25] and ours are supervised by image-level labels, while [13] is supervised by both image-level labels and partially bounding box-level annotations.

lesion region localization significantly exceeds [13] under all the thresholds. Besides, the performance of some pathologies, such as "Mass", "Pneumonia", and "Pneumothorax" are very close to [13]. But we also notice that the performance of "Atelectasis" is inferior to [25]. And for "Cardiomegaly", the localization accuracy is lower than [25] when T(IoU) is less than 0.3, while it is slightly higher than [25] 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 [25] under the same setting, which serves as an explanation of our superior performance.

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 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 regions on the heatmaps in Fig. 5.

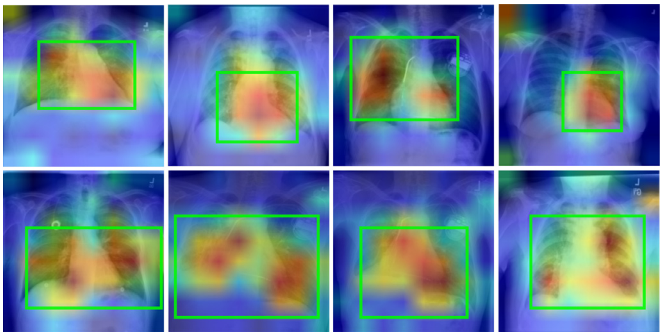


Fig. 5. Examples of heatmaps for "no finding" images. The cropped regions are denoted by green bounding boxes. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

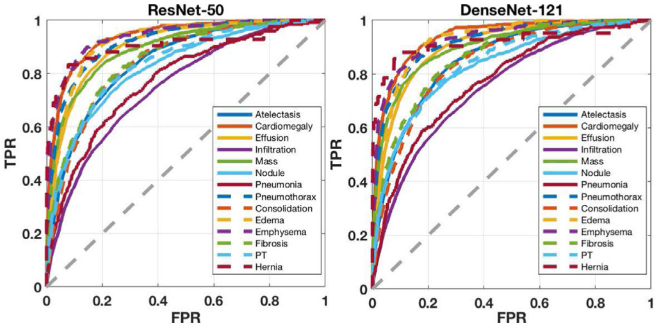


Fig. 6. ROC curves of AG-CNN on the 14 diseases (ResNet-50 and DenseNet-121 as backbones, respectively).

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. 6. 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.

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 4.1 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 [12,19,25,29] on the ChestX-ray14 dataset. Wang et al. [25] classify and localize the thorax dis-

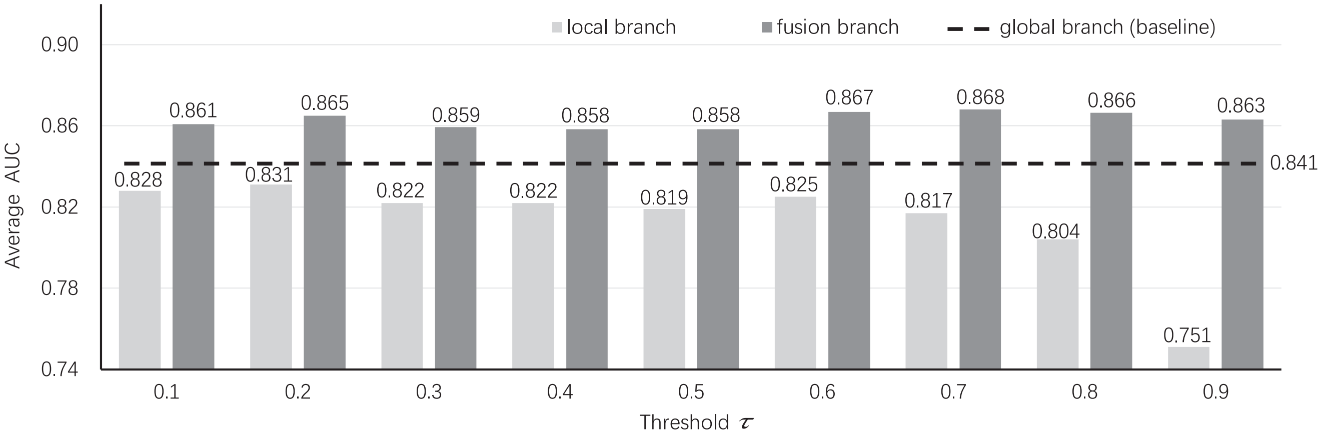


Fig. 7. Average AUCs for different settings of τ on the test set (ResNet-50 as backbone). Note that the results from global branch are our baseline.

Table 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

Table 4
Results corresponding different statistics.

Statistic	Global	Local	Fusion
Max	0.8412	0.8171	0.8680
L1	0.8412	0.8210	0.8681
L2	0.8412	0.8213	0.8672

ease in a unified weakly supervised framework. The reported results from Yao *et al.* [29] are based on the model in which labels are considered independent. Kumar et al. [12] 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 [19] by 2.9%. AUC scores of pathologies such as *Cardiomegaly* and *Infltration* are higher than [19] by about 0.03. AUC scores of *Mass, Fibrosis* and *Consolidation* surpass [19] by about 0.05. Furthermore, we train AG-CNN with 70% of the dataset, but 80% are used in [12,19]. In nearly all the 14 classes, our method yields best performance. Only Rajpurkar *et al.* [19] 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 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 4.1. We present the classification performance of these training strategies in Table 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 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, we observe that a better performance in local branch leads to better performance in fusion branch. We infer that the performance of local branch is essential to enhance the whole framework.

Variant of heatmap analysis. In Table 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.

4.3. Parameter analysis

We analyze the sensitivity of AG-CNN to the parameter consists in τ in Eq. 4, which defines the local region and affects the classification accuracy. Fig. 8 shows the average 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. 7 compares the average AUC

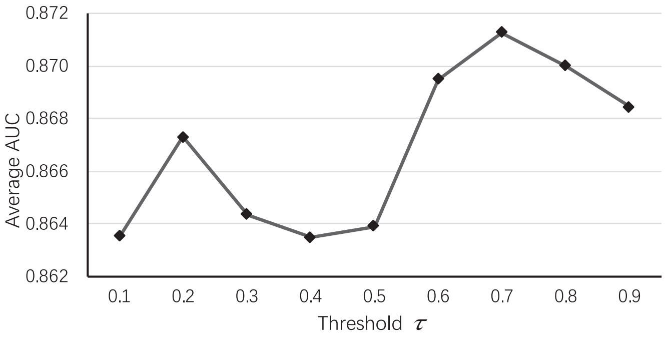


Fig. 8. Average AUC scores of AG-CNN with different settings of τ on the validation set (ResNet-50 as backbone).

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].

5. Conclusion

In this paper, we propose an attention guided convolutional neural network for thorax disease 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. In the future research, we will continue to investigate more accurate lesion localization method to improve the recognition performance.

Declaration of Competing Interest

We wish to draw the attention of the Editor to the following facts which may be considered as potential conflicts of interest and to significant financial contributions to this work. We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome. We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us. We confirm that we have given due consideration to the protection of intellectual property associated with this work and that there are no impediments to publication, including the timing of publication, with respect to intellectual property. In so doing we confirm that we have followed the regulations of our institutions concerning intellectual property. We understand that the Corresponding Author is the sole contact for the Editorial process (including Editorial Manager and direct communications with the office). He/she is responsible for communicating with the other authors about progress, submissions of revisions and final approval of proofs. We confirm that we have provided a current, correct email address which is accessible by the Corresponding Author and which has been configured to accept email from yphuang@bjtu.edu.cn.

Acknowledgment

This work was supported by the Fundamental Research Funds for the Central Universities (2019JBZ104).

References

[1] M. Akbari, L. Nie, et. al., AMMAmm: towards adaptive ranking of multi-modal documents, IJMIR 4 (4) (2015) 233–245.
[2] J. Deng, W. Dong, et. al., Imagenet: A large-scale hierarchical image database, in: Proceedings of the CVPR, 2009, pp. 248–255.
[3] H. Fu, Y. Xu, et. al., Segmentation and quantification for angle-closure glaucoma assessment in anterior segment oct, IEEE TMI 36 (9) (2017) 1930–1938.
[4] Q. Guan, Y. Huang, Multi-label chest x-ray image classification via category-wise residual attention learning, Proceedings of the PRL (2018).
[5] S. Guendel, S. Grbic, et al., Learning to recognize abnormalities in chest x-rays with location-aware dense networks, Iberoamerican Congress on Pattern Recognition (2018) 757–765.
[6] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the CVPR, 2016, pp. 770–778.
[7] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[8] G. Huang, Z. Liu, et al., Densely connected convolutional networks, Proceedings of the CVPR (2017) 4700–4708.
[9] Y.-G. Jiang, C.-W. Ngo, et. al., Towards optimal bag-of-features for object categorization and semantic video retrieval, in: Proceedings of the ACM ICIVR, ACM, 2007, pp. 494–501.
[10] A. Krizhevsky, I. Sutskever, et. al., Imagenet classification with deep convolutional neural networks, in: Proceedings of the NIPS, 2012, pp. 1097–1105.
[11] A. Kumar, J. Kim, et. al., An ensemble of fine-tuned convolutional neural networks for medical image classification, IJBHI 21 (1) (2017) 31–40.
[12] P. Kumar, M. Grewal, et al., Boosted cascaded convnets for multilabel classification of thoracic diseases in chest radiographs, International Conference Image Analysis and Recognition (2018) 546–552.
[13] Z. Li, C. Wang, et al., Thoracic disease identification and localization with limited supervision, Proceedings of the CVPR (2018) 8290–8299.
[14] G. Litjens, T. Kooi, et al., A survey on deep learning in medical image analysis, Medical Image Analysis 42 (2016) 60–88.
[15] Z. Luo, A. Mishra, et. al., Non-local deep features for salient object detection, in: Proceedings of the CVPR, 2017, pp. 6609–6617.
[16] A. Melo, H. Paulheim, Local and global feature selection for multilabel classification with binary relevance, AI Reviews 51 (1) (2019) 33–60.
[17] K. Murphy, A. Torralba, et. al., Object detection and localization using local and global features, in: Toward Category-Level Object Recognition, Springer, 2006, pp. 382–400.
[18] A. Qayyum, S.M. Anwar, et al., Medical image analysis using convolutional neural networks: a review, Journal of medical systems 42 (11) (2018) 226.
[19] P. Rajpurkar, J. Irvin, et al., Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists, PLoS Medicine 15 (11) (2018) e1002686.
[20] R. Roy, T. Chakraborti, A.S. Chowdhury, A deep learning-shape driven level set synergism for pulmonary nodule segmentation, Proceedings of the PRL (2019).
[21] C.-R. Shyu, C. Brodley, et. al., Local versus global features for content-based image retrieval, in: Proceedings of the IEEE Workshop on CAIVL, 1998, pp. 30–34.
[22] K. Simonyan, A. Zisserman, in: Very deep convolutional networks for large-scale image recognition, 2015.
[23] C. Szegedy, W. Liu, et. al., Going deeper with convolutions, in: Proceedings of the CVPR, 2015, pp. 1–9.
[24] Y. Tang, X. Wang, et. al., Attention-guided curriculum learning for weakly supervised classification and localization of thoracic diseases on chest radiographs, in: Proceedings of the IWMLMI, Springer, 2018, pp. 249–258.
[25] X. Wang, Y. Peng, et. al., Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in: Proceedings of the CVPR, IEEE, 2017, pp. 3462–3471.
[26] X. Wang, Y. Peng, et. al., Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays, in: Proceedings of the CVPR, 2018, pp. 9049–9058.
[27] Y. Wu, Y. Lin, et al., Progressive learning for person re-identification with one example, IEEE TIP 28 (6) (2019) 2872–2881.
[28] Z. Yang, T. Luo, et. al., Learning to navigate for fine-grained classification, in: Proceedings of the ECCV, 2018, pp. 420–435.
[29] L. Yao, E. Poblenz, et. al., Learning to diagnose from scratch by exploiting dependency among labels, arXiv:1710.10501 (2017).
[30] Z. Zhong, L. Zheng, et al., Camstyle: a novel data augmentation method for person re-identification, IEEE TIP 28 (3) (2019) 1176–1190.
[31] L. Zhu, Z. Xu, et al., Uncovering the temporal context for video question answering, IJCV 124 (3) (2017) 409–421.
[32] J. Zou, W. Li, et. al., Scene classification using local and global features with collaborative representation fusion, Inf. Sci. 348 (2016) 209–226.