MULTI-ATTENTION NETWORK FOR THORACIC DISEASE CLASSIFICATION AND LOCALIZATION

Yanbo Ma *Qiuhao Zhou* *Xuesong Chen* *Haihua Lu* *Yong Zhao**

School of Electronic and Computer Engineering, Shenzhen Graduate School of Peking University
Shenzhen, China

ABSTRACT

The chest X-ray is one of the most commonly available radiological examinations for diagnosing lung diseases. This task remains a major challenge due to 1) the shortage of accurate annotations for chest X-ray examinations, 2) the diversity of lesion areas on X-rays from different thoracic disease and 3) the problem of class imbalance in existing chest X-ray databases. In this paper, we propose a new multi-attention convolutional neural network for thoracic disease classification and localization. First, the framework is equipped with squeeze-and-excitation (SE) block as a feature attention module to offer a chance of cross-channel feature recalibration. Second, we propose a novel space attention module to combine global and local information. Third, we present a hard examples attention module to alleviate the class imbalance problem. The comprehensive experiments are performed on the ChestX-ray14 dataset. Quantitative and qualitative results demonstrate that our method outperforms the state-of-the-art algorithm.

Index Terms— Chest X-ray, attention mechanism, convolution neural network, medical image processing

1. INTRODUCTION

More than 1 million adults are hospitalized with pneumonia and around 50,000 patients die from the disease every year in the US alone [1].Generally, most of X-rays mainly rely on radiologists' manual observation. It is time-consuming and requires a high degree of skill and concentration. Due to the complexity of chest radiographs and the shortage of expert radiologists, automatic chest X-ray image classification and localization is becoming an increasingly important technique to support the clinical diagnosis of thorax diseases.

Several existing works on chest X-rays classification typically employ the global image for training. For example, [2] presents a unified weakly-supervised multi-label image classification and pathology localization framework. He performs the experiment on the pre-trained models (using ImageNet [3], e.g., AlexNet [4], GoogLeNet [5], VGGNet-16 [6] and ResNet-50 [7]). However, most of chest X-ray images possess multiple class labels and the same lesion area may have more than one thoracic disease. Those previous works fail to treat different diseases separately and the classifier could be confused when detecting a certain type of disease by other diseases' features. Another important issue is that the lesion area is extremely small compared with the global image and their position is not fixed. Researchers generally use the ChestX-ray14 dataset which has 14 diseases' X-ray images, but each specific disease possesses a small proportion in the dataset. Therefore, the problem of class imbalance also stall the advancement of automatic chest X-ray diagnosis.

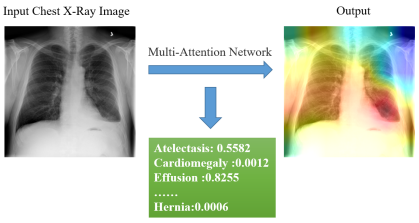


Fig. 1. An example of our multi-attention convolutional neural network for thoracic disease diagnosis. The input is a chest X-ray image and the output is prediction scores and localization heat map for the diseases.

In this paper, our network can classify thoracic diseases precisely and localize the disease regions on X-rays at pixel-level granularity. Fig.1 demonstrates an example concerning the output of our model. The multi-attention network is featured in three aspects. First, we use the squeeze-and-excitation (SE) attention module [10] which can adaptively recalibrates channel-wise feature responses and reinforces the sensitivity of our model by explicitly modeling interdependencies between channels. Second, to focus on the disease-specific local features, we propose the global and local attention module in the feature fusion layer. Third, we present the two-stage training method to alleviate the problem of class imbalance. Our quantitative results show that our multi-attention model achieves significant accuracy improvement on disease identification.

*Corresponding author

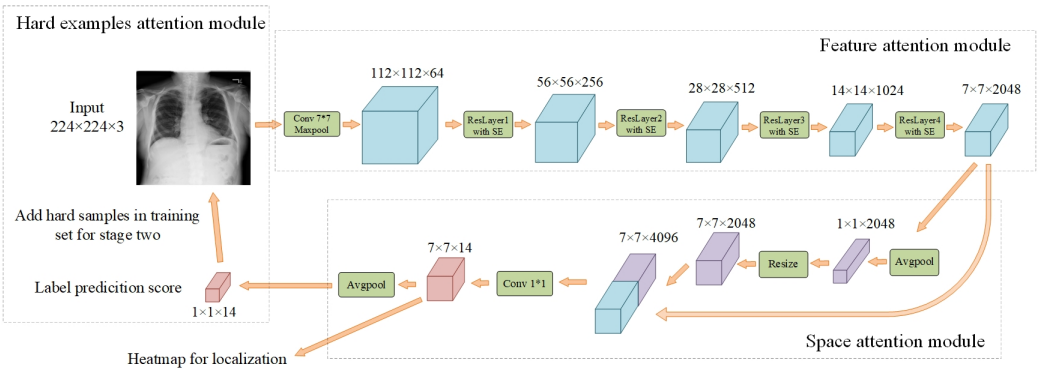


Fig. 2. The overall flowchart of our multi-attention network and disease localization process.

2. RELATED WORK

Recent surveys [11, 12, 13] have demonstrated that deep learning technologies have been extensively applied to the field of chest X-ray image annotation, classification, and localization. The multi-label classification problem associates each instance with a subset of possible labels. The simplest approach is to break the multi-label classification problem into independent binary classification problems. The random k-labELsets (RAKEL) algorithm [14] can construct each member of the ensemble by considering a small random subset of labels and learning a single-label classifier for the prediction of each element in the powerset of this subset. [2] trains a multi-label convolutional neural networks(CNNs) classification model of thoracic diseases based on the pre-trained models. [8] exploits the conditional dependencies among abnormality labels for better diagnostic results. [15] puts forward a weakly supervised deep learning framework equipped with SE-blocks, multi-map transfer, and max-min pooling for classifying thoracic diseases as well as localizing suspicious lesion regions. Different from the previous global learning methods, we make use of the multi-attention mechanism and fuse the local and global information for enhancing the classification performance.

3. PROPOSED APPROACH

In this section, the technical details of the proposed multi-attention network will be explicitly described. The illustration of the proposed architecture is summarized in Fig. 2:

3.1. Feature Attention Module

The feature attention module follows the structure of ResNet-101 [7]. However, this task is different from the traditional multi-classification problem because abnormal patterns are usually featured with complex interactions. For example, a patient who has cardiomegaly is more likely to additionally have pulmonary edema. Classical CNNs are initialized and trained independently so that they cannot recognize the potential for these interdependencies.

It has been proved that SE-block [10] can explicitly model channel-interdependencies within modules. Inspired by it, we insert SE-block into every ResNet block to make sure that the network is able to increase its sensitivity to useful informative features. SE-block contains two steps: squeeze and excitation. A diagram of an SE-block is shown in Fig. 3

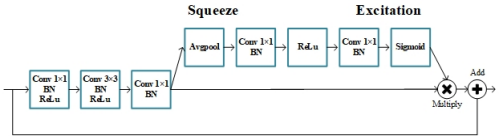


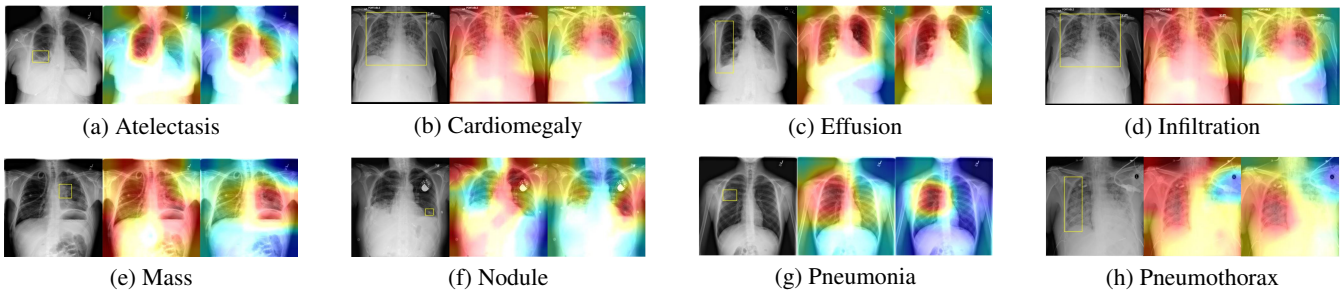
Fig. 3. Illustration of an SE-Block in ResNet block.

Squeeze. We use global average pooling to squeeze the C feature maps after the convolution into a feature vector of C length for exploiting channel dependencies. This operation conducts feature compression along the spatial dimension and the two-dimensional feature channel turns into a real number which has a global receptive field. It represents the global distribution of responses on feature channels and allows the layer near the input to obtain the global sensory field, which is extremely useful for many tasks. Formally, u is the transformed output after the convolution and $H \times W$ is the spatial dimension. The c-th element of z, the vector after squeezing, is calculated by:

$$z^{c} = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u^{c}(i,j) \tag{1}$$

Excitation. This operation is to fully capture channel-wise dependencies. The function should learn both nonlinear interactions between channels and non-mutually-exclusive relationships, and therefore we use a gating mechanism with a sigmoid activation:

$$s = \sigma(W_2 \times ReLU(W_1 z)) \tag{2}$$



(a) Atelectasis (b) Cardiomegaly (c) Effusion (d) Infiltration

(e) Mass (f) Nodule (g) Pneumonia (h) Pneumothorax

Fig. 4. Examples of heatmaps for localization. In each pair, the left image is the original chest X-ray image, the middle one is the localization result from the model with the average pooling operation between the ResNet5c and classifiers and the right is the localization result from the model with the average pooling operation in the end.

where s is the channel-wise attention coefficient for feature recalibration, W_1 and W_2 are the weight of convolution, σ is a sigmoid activation function. For channel c, we multiply channel-wise attention coefficient s and the transformed feature maps.

3.2. Space Attention Module

The outputs of the feature attention module are feature maps with the size of 7×7. Generally, researchers use fully-connected layers to get the results. However, this method meets two problems. First, the classifier will receive lots of noise when the network converts feature maps to feature vectors by pooling operation for the interest of areas that only occupy the small space of feature maps. Second, the classifier only learns from the global information and it can't locate the disease precisely. To handle these problems, the classifiers should pay attention to the local information.

Inspired by [16], we propose the space attention module as shown in Fig. 2. First, we get the feature map of ResNet-101. Then we apply global average pooling operation on it. Here, we get the feature vector of this image. After that, we resize it to 7×7 and concatenate it to the original features. From this module, each pixel of the feature map has global information and the classifier gets the largest receptive field. Therefore, the classifier can get the global and local information whether for localization and classification.

Notably, the convolution operations like [17] are not conducted in our network. We only perform global average pooling operation and just resize it to its original shape instead of introducing convolution operation between pooling and resizing operation. Due to the lack of samples, a large number of parameters will bring about the problem of overfitting and make the network hard to be trained. Therefore, we only make the global branch just like the common classification network. After this module, the global branch provides all the underlying information about image while the local branch makes it possible to pay attention to lesion areas and alleviate the disturbed noisy of other areas.

3.3. Hard example attention module

The prediction of the model will be biased towards a certain category if there is a large gap between the proportion of positive and negative samples. Our task is to distinguish 14 diseases in ChestX-ray14 dataset. For each disease, the mount of negative samples is several times more than positive samples. In order to solve this problem, we propose a two-stage training method.

In the first stage, our main task is to find hard positive cases. We use the original training set to train the network. The original training set is denoted as set **A**. We use the trained model to select the training set and each image will receive a predicted score of each disease. We set the threshold as 0.5. If the predicted score of a disease is greater than 0.5, we will consider that this sample has the corresponding disease. Set **B** consists of the positive samples whose prediction is different from the label. In other words, the elements in set **B** are hard positive samples. In the second stage, we combine set **B** with set **A** to form a new set **C**, where the proportion of positve examples is increased. We retrain the network by using the elements in set **C** to distinguish the 14 diseases.

Different from the general resampling method, we not only increase the proportion of positive samples, but also increase the proportion of hard positive samples which have more information.

4. EXPERIMENTS

4.1. Dataset and Experimental Settings

We conduct experiments on the ChestX-ray14 dataset [2] which consists of 112120 frontal-view chest X-ray images of 30805 unique patients with 14 disease labels (each image can have multi-labels).The dataset is extracted from the clinical PACS database at the National Institutes of Health Clinical Center. Till now, it is the largest public chest X-rays dataset. These chest X-ray images are originally released in the PNG format and are rescaled to the size of 1024×1024. We use the same patient-level data splits provided in the dataset officially

[2], which uses roughly 70% of the images for training, 10% for validation and 20% for testing.

In training, we perform data augmentation to prevent overfitting and increase the data volume. First, we downscale the images to 256×256. Second, we adopt random translation from -12 pixels to 12 pixels and our crop size is 224×224. Third, we normalize the image based on the mean and standard deviation of images in the ImageNet training set. The model is initialized using weights from the pre-trained ResNet-101 model(using ImageNet). We optimize the network using SGD, the weight decay of 0.0001 and the momentum of 0.9. Then, the batch size is set to 96. The learning rate starts from 0.01 and is divided by 10 after 20 epochs.

4.2. Results

Our result is shown in Table.1 and the ROC curve is shown in Fig.5. Besides, we compare the classification performance of our multi-attention network to the state-of-the-art deep learning methods. Our proposed model achieves the highest average per-class AUC score of 0.7941 and outperforms in classification of 11 out of the 14 diseases.

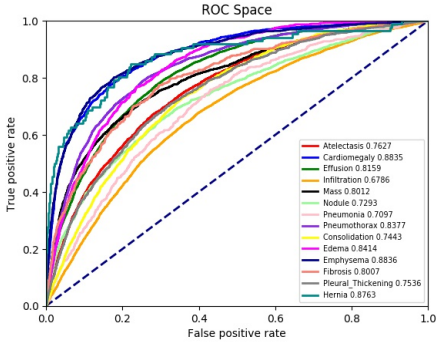


Fig. 5. The ROC curve of multi-attention network on the 14 diseases.

We use the heatmap to visualize the most indicative pathology areas on X-rays for the localization of lesion regions. Heatmaps are computed from the model without the last average pooling. Then, we compare our results with the X-ray image which is annotated with lesion regions by radiologists. The heat map can intuitively locate the disease regions without any accurate annotation. Some examples are shown in Fig.4.

4.3. Ablation study

As shown in Table.1, we conduct ablation experiments to prove the effectiveness of our proposed attention modules. We first test the performance of original ResNet architecture and the average AUC score is 0.7474 for all 14 diseases. To highlight some features, we introduce the SE module into each block of ResNet. It can be found that the average AUC score rises up 1.6% when the feature attention module is added to the network. We continue to test the performance of the space attention module and the 1.2% lift demonstrates the effect from concatenating the global and local information. In addition to that, hard true examples need gain more attention from the network. The hard examples attention module improves the average AUC from 0.7758 to 0.7941. It also shows that improving the ratio of hard examples in the total example and adjusting the distribution of data according to the hard example are also vital for training this X-ray dataset.

Table 1. The multi-label classification AUC score comparison on the test set of ChestX-ray14. BSL is baseline model(ResNet-101). FAM is the feature attention mudule. SAM is the space attention module. MA is multi-attention module which includes the feature attention mudule, the space attention module and the hard examples attetion module.

Thorax Disease	[2]	[18]	BSL	BSL +FAM	BSL +FAM +SAM	MA
Atelectasis	0.7003	0.7320	0.7408	0.7432	0.7581	**0.7627**
Cardiomegaly	0.8100	0.8440	0.8566	0.8829	0.8784	**0.8835**
Effusion	0.7585	0.7930	0.8063	**0.8184**	0.8138	0.8159
Infltration	0.6614	0.6660	0.6830	**0.6916**	0.6706	0.6786
Mass	0.6933	0.7250	0.7262	0.7730	0.7928	**0.8012**
Nodule	0.6687	0.6850	0.6651	0.7041	0.7274	**0.7293**
Pneumonia	0.6580	**0.7200**	0.6751	0.6848	0.6854	0.7097
Pneumothorax	0.7993	**0.8470**	0.8076	0.8143	0.8276	0.8377
Consolidation	0.7032	0.7010	0.7225	0.7367	0.7369	**0.7443**
Edema	0.8052	0.8290	0.8197	0.8254	0.8296	**0.8414**
Emphysema	0.8330	0.8650	0.8226	0.8408	0.8665	**0.8836**
Fibrosis	0.7859	0.7960	0.7643	0.7478	0.7911	**0.8007**
Pleural Thickenin	0.6835	0.7530	0.7256	0.7351	0.7389	**0.7536**
Hernia	0.8717	0.8760	0.6479	0.6917	0.7436	**0.8763**
Average	0.7451	0.7724	0.7474	0.7636	0.7758	**0.7941**

5. CONCLUSION

In this paper, we propose a multi-attention network for thorax disease classification and localization. The proposed network consists of three attention modules: feature attention module for cross-channel feature recalibration, space attention module for including both global and local information and hard example attention module for alleviating class imbalance problem. In addition, our network can also provide heatmaps to assist doctors determine the location of lesions. Both quantitative and qualitative results have indicated that our framework outperformed the state-of-the-art result.

6. ACKNOWLEDGEMENTS

This work was supported by Science and Technology Planning Project of Shenzhen(No. NJYJ20170306091531561), Science and Technology Planning Project of Shenzhen(No. JCYJ2016050617265 1253), and National Science and Technology Support Plan, China(No. 2015BAKO1B04).

7. REFERENCES

[1] "Pneumonia can be preventedłvaccines can help," https://www.cdc.gov/Features/Pneumonia/, 2017.

[2] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, 2017, pp. 3462–3471.

[3] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[5] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[6] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[8] Li Yao, Eric Poblenz, Dmitry Dagunts, Ben Covington, Devon Bernard, and Kevin Lyman, "Learning to diagnose from scratch by exploiting dependencies among labels," *arXiv preprint arXiv:1710.10501*, 2017.

[9] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger, "Densely connected convolutional networks.," in *CVPR*, 2017, vol. 1, p. 3.

[10] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," *arXiv preprint arXiv:1709.01507*, vol. 7, 2017.

[11] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen AWM van der Laak, Bram Van Ginneken, and Clara I Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[12] Dinggang Shen, Guorong Wu, and Heung-Il Suk, "Deep learning in medical image analysis," *Annual review of biomedical engineering*, vol. 19, pp. 221–248, 2017.

[13] Shin Hoo-Chang, Holger R Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M Summers, "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1285, 2016.

[14] Zhe Li, Chong Wang, Mei Han, Yuan Xue, Wei Wei, Li-Jia Li, and F Li, "Thoracic disease identification and localization with limited supervision," *arXiv preprint arXiv:1711.06373*, 2017.

[15] Chaochao Yan, Jiawen Yao, Ruoyu Li, Zheng Xu, and Junzhou Huang, "Weakly supervised deep learning for thoracic disease classification and localization on chest x-rays," in *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. ACM, 2018, pp. 103–110.

[16] Wei Liu, Andrew Rabinovich, and Alexander C Berg, "Parsenet: Looking wider to see better," *arXiv preprint arXiv:1506.04579*, 2015.

[17] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia, "Pyramid scene parsing network," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2881–2890.

[18] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M Summers, "Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, 2018, pp. 9049–9058.