Weakly Supervised Deep Learning for Thoracic Disease Classification and Localization on Chest X-rays

Chaochao Yan
The University of Texas at Arlington
chaochao.yan@mavs.uta.edu

Jiawen Yao
The University of Texas at Arlington
jiawen.yao@mavs.uta.edu

Ruoyu Li
The University of Texas at Arlington
ruoyu.li@mavs.uta.edu

Zheng Xu
The University of Texas at Arlington
zheng.xu@mavs.uta.edu

Junzhou Huang*
The University of Texas at Arlington
Tencent AI Lab
jzhuang@uta.edu

ABSTRACT

Chest X-rays is one of the most commonly available and affordable radiological examinations in clinical practice. While, detecting thoracic diseases on chest X-rays is still a challenging task for machine intelligence, due to 1) the highly varied appearance of lesion areas on X-rays from patients of different thoracic disease and 2) the shortage of accurate pixel-level annotations by radiologists for model training. Existing machine learning methods are unable to deal with the challenge that thoracic diseases usually happen in localized disease specific areas. In this article, we propose a weakly supervised deep learning framework equipped with squeeze-and-excitation blocks, multi-map transfer and max-min pooling for classifying common thoracic diseases as well as localizing suspicious lesion regions on chest X-rays. The comprehensive experiments and discussions are performed on ChestX-ray14 dataset. Both numerical and visual results have demonstrated the effectiveness of proposed model and its better performance against the state-of-the-art pipelines.

CCS CONCEPTS

• **Theory of computation → Machine learning theory**; • **Applied computing →** *Imaging*;

KEYWORDS

Chest X-ray, computer-aided diagnosis, weakly-supervised learning, fully convolutional network, multi-map transfer layer, feature recalibration

1 INTRODUCTION

Chest X-ray imaging is currently one of the most widely available radiological examinations for screening and clinical diagnosis. However, automatic understanding of chest X-ray images is currently a technically challenging task due to the complex pathologies of different sorts of lung lesions on images. In clinical practice, the analysis and diagnosis based on chest X-rays are heavily dependent on the expertise of radiologists with at least years of professional experience. Therefore, there is a critical need of a computer-aided system that is able to automatically detect different types of thoracic diseases merely from reading patients' chest X-ray images. This is all founded on a well-designed transfer of human knowledge to machine intelligence.

Since the last decade of years, as a promising technology, Medical Artificial Intelligence (Medical AI) has globally attracted interest. Especially after the emergence and fast progress of deep learning, a revolution of computer-aided diagnosis (CAD) technique has officially started and impacted in many bio-medical applications, e.g. diabetic eye disease diagnosis [11], cancer metastases detection and localization [3, 20, 23], lung nodule detection [27], and survival analysis [37], etc. However, introducing deep learning as solution to reading and understanding chest X-ray images is challenging due to the following reasons: 1) the visual patterns extracted from samples of different types of thoracic diseases are usually highly diverse in their appearance, sizes and locations (examples of common thoracic diseases in ChestX-ray14 dataset [33] are available in Fig.1) ; 2) retrieving massive high-quality annotations of disease, such as focal zone, on chest X-ray images is not affordable. The expenses result from both the cost of hiring experienced radiologists and the hardware requirements of collection, storage, processing of those data. Therefore, ChestX-ray14, although as the largest and most quality public chest X-rays dataset, does not provide with any pixel-wise annotations or coarse bounding boxes (example of which is in Fig.1) for most of chest X-ray images. Consequently, it is obvious that any machine learning models proposed to be compatible with ChestX-ray14 dataset are required to work merely with image-level class label plus a very small amount of bounding box annotations.

Many research efforts have been made for automatic detection of thoracic diseases based on diverse data generated by chest X-ray scanning. Chapman et al. [2] discussed the performance of Bayesian network and decision tree at identifying chest X-ray reports. Ye et al. [36] reduced false positive in classification of lung

*This work was partially supported by US National Science Foundation IIS-1423056, CMMI-1434401, CNS-1405985, IIS-1718853 and the NSF CAREER grant IIS-1553687.

ACM-BCB'18, August 29-September 1, 2018, Washington, DC, USA

ACM ISBN 978-1-4503-5794-4/18/08…$15.00
https://doi.org/10.1145/3233547.3233573

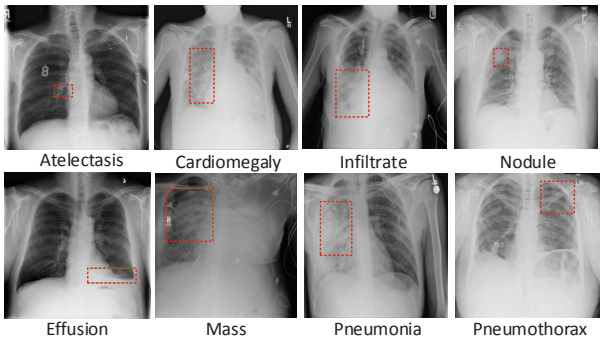

Figure 1: Examples of chest X-rays of eight thoracic diseases and associated lesion regions from ChestX-ray14 [33]. The regions were annotated as red bounding boxes by radiologists. The bounding boxes were only used for evaluation.

nodules on chest X-rays via introducing a weighted support vector machine (SVM) classifier. Beyond hand-crafted features, Wang et al. [33] concatenated the classifier to a fully convolutional network (FCN) as feature extractor in classification of thoracic diseases on images from ChestX-ray14 dataset, in which they compared four classic convolutional neural network (CNN) architectures, i.e., AlexNet [18], VGGNet [29], GoogLeNet [30], ResNet [12]. Later, Yao et al. [34] investigated the hidden correlations among the 14 pathological class labels in ChestX-ray14 dataset. The most recent framework proposed by Rajpurkar et al. is CheXNet [25] that fine-tuned a revised 121-layer DenseNet [14] on ChestX-ray14 images and achieved the state-of-the-art performance on thoracic disease detection. However, those previous works typically employ single or multiple fully connected layers to densely connect and select significant features on the feature maps generated by convolutional networks. As a consequence, this architecture and its similar variants do not treat different diseases separately and thus ignore a crucial fact that those lesion areas on chest X-rays actually are disease specific. Another important issue is that many images from ChestX-ray14 contain lesion areas of more than one thoracic disease (i.e. most of images have multiple class labels). This setting is to simulate a common case that a radiologist often deals with in clinical practice. Intuitively, in their models, the classifier could be possibly confused when detecting a certain type of disease by those features extracted from the lesions belong to other diseases. Therefore, a significant improvement is expected through learning disease-discriminative features on chest X-rays.

In this paper, we will present a novel weakly-supervised learning model to particularly overcome the aforementioned issues existing in previous works. The proposed model is able to classify thoracic diseases merely reading provided chest X-rays as well as to localize the disease regions on X-rays at pixel-level granularity. First, we harnessed the latest Fully Convolutional Network (FCN) alike model, i.e. DenseNet [14], as backbone network, because DenseNet has obviously shown its outstanding performance on generic image classification [14] and semantic segmentation [16]. Much beyond the original DenseNet, for the first time, we proposed to use the so-called "Squeeze-and-Excitation" (SE) block [13], which aims to reinforce the sensitivity of our model to subtle differences between normal and lesion regions by explicitly modeling the channel interdependencies. Moreover, we incorporated the use of multi-map transfer layers to make our network perform better to learn disease-specific features that are highly related to disease modalities, e.g. "Atelectasis" and "Nodule" on chest X-ray. The last but not the least, we realized that the max-min pooling operators [7] perform better at spatially squeezing feature maps for each class of disease. Our major contributions in the paper are summarized as follows:

- A "Squeeze-and-Excitation" block was embedded after convolution layer in DenseNet block for feature recalibration.
- Concatenate stacked multi-map transfer layers to DenseNet replacing fully connected layers to mitigate the multi-label issue, which becomes crucial when labels are noisy.
- We incorporated the max-min pooling operator to aggregate spatial activations from multi-maps into a final prediction.
- Extensive experiments have been conducted to demonstrate the effectiveness of proposed methods. Our method achieved superior performance compared to the state-of-the-arts.
- The effectiveness of each proposed component are individually verified by experiments on ChestX-ray14 dataset.

The rest of the paper is organized as follows. Problem description and recent work on automatic detection and diagnosis techniques on chest X-rays were given in Section 1&2. Then, we presented our framework details in Section 3. Experimental setup, results and discussions were in Section 4. In last section, we summarized our contributions as well as the future research highlights.

2 RELATED WORK

For a long time, designing a computer-aided diagnosis platform to understand radiographs has widely attracted research interest. A well-prepared database is one of the most significant factors in successfully developing a generalizable machine learning model, especially a data-hungry deep neural network model. JSRT released a chest X-ray image set [28] which contains 247 chest X-ray images including 93 normal images and 154 of those exhibited malignant and benign lung nodules. Due to the limited size of JSRT data, it is difficult to train a complex model against over-fitting. [8] trained a convolutional network based classifier for lung nodules classification on JSRT dataset and its improved version BSE-JSRT dataset [32] in which bone shadows were excluded. The Indiana chest X-ray dataset [5] has a mixed collection of 8,121 frontal and lateral view X-ray images together with 3,996 radiology reports contain labels from trained experts. [15] compared the performance of multiple state-of-the-art deep learning models on Indiana dataset for disease classification and localization of remarkable regions that contribute most to an accurate classification.

In [33], a hospital-scale database, ChestX-ray14, that comprises 108,948 frontal-view of X-ray images of 32,717 individual patients was presented together with the 14 classes of image labels, each of which corresponds to a thoracic disease. Therefore, each image may have multiple labels. ChestX-ray14 is probably the largest, most quality, X-ray image dataset available publicly. It is notable that the aforementioned image labels are not directly from manual annotation by pathologists, for instead, were mined by natural language processing technique [1, 19] on associated radiaological reports.

Consequently, the class labels in training set is noisy, which brings extra challenge to disease classification task. Besides, [33] experimentally demonstrated that those common thoracic diseases could be correctly detected or even spatial-localized via a unified weakly-supervised multi-label learning framework trained by generated noisy weak class labels. The ResNet outperformed other popular convolutional neural networks, e.g. AlexNet [18], GoogLeNet [30] and VGGNet-16 [29] by rendering class-wise ROC-AUC scores e.g. 0.8141 for "Cardiomegaly". While, for some diseases like "Mass" and "Pneumonia", the scores were dramatically dragged to 0.5609 and 0.6333 respectively. This result disclosed the long-standing ignorance of the incapability of traditional CNNs on learning meaningful representations with weak supervision of noisy labels. However, the major difficulty of applying deep learning models on medical problems is the shortage of high-quality annotations by pathologist.

Shortly after ChestX-ray14 was released, [25] proposed a state-of-the-art CNN model named as CheXNet that consists of 121 layers. The model accepts chest X-ray images as input and outputs the probability of disease along with a heatmap which localizes the most indicative regions of disease on the input images. On the task of detecting pneumonia, the CheXNet successfully exceeded the average performance of four experienced radiologists on a subset of 420 X-ray images of pneumonia patients. However, the network in [25] is a variant of DenseNet [14] without any significant modifications particularly for learning representations under a weak supervision. The network was initialized by weights pretrained on ImageNet [6], the content of which shares few in common with the images of ChestX-ray14. The lower-level representations learned on ImageNet are not guaranteed to accurately customize the shape and the contour of regions of thoracic diseases. Even though [25] has lifted the classification accuracy by a margin of 0.05 on ROC-AUC score, it still left quite a space for improvement.

As mentioned in [25], the significance of comparison between CheXNet and human pathologist labeling was compromised by the fact that only the frontal view of radiographs were presented to pathologists, and it has been confirmed that there are 15% successful diagnoses of pneumonia by pathologists mainly contributed by the lateral views, which were not available in ChestX-ray14. Consequently, a multi-view version of chest X-ray dataset - *MIMIC-CXR* was presented in [26], and based on which a dual deep convolutional network framework was naturally proposed to utilize both frontal and lateral views, if given, for disease classification. While, the network for each view was separately trained instead of weights sharing. The outputs of each network (view) were concatenated as a unified vector before a set of final fully connected layers for generating multi-class prediction. However, because of the lack of other view of radiographs in ChestX-ray14, [26] did not include a face-to-face comparison with CheXNet on the same dataset. Therefore, the actual effectiveness of introducing another relevant view of X-ray is . Moreover, the numerical results of [26] has not strongly supported the conclusion that combining more views of radiographs brings lift on recognition performance without learning the correlation between views.

As discussed above, training a classifier on X-ray images is more difficult than generic image, e.g. ImageNet, where object of interest is usually positioned in the middle of image. The lesion area of lung could be pretty small compared to the entire X-ray images. Besides, the variant condition of capturing, e.g. posture of patient, brings extra distortion and misalignment. To address these problems, [9] proposed an attention guided convolutional neural network (AG-CNN) to extract regions of interest (RoI) as a rough localization of lesion areas from the last convolution outputs of global network which train on raw X-ray images with class label supervision. Then, extracted RoI patches were fed to a separate local branch of CNN for learning local representation of lesion. At last, a fusion branch concatenates features generated by both global and local branches with a fine-tune with several fully-connected layers.

The ChestX-ray14 offers a very noisy class-labels and quite a few bounding boxes as ground-truth for regions of interest localization. This makes it a classic weakly supervised learning problem [4, 31], which is pretty common in medical areas and becoming important when developing AI in fields where expertise is expense. [35] modeled the problem as multiple instance learning (MIL) on X-ray as a roughly-labeled bag of patches. They parameterised the Log-Sum-Exp pooling with a trainable lower-bounded adaptation (LSE-LBA) to construct illustrative saliency map at multiple resolutions.

3 METHODOLOGY

In this section, we will explicitly present the technical details of proposed framework. First, we illustratively discuss the advantages of DenseNet compared with other modern FCN models. Then, we individually discuss the roles of the three components that bring extra performance lift beyond DenseNet: squeeze-and-excitation block, multi-map transfer layer and max-min pooling operator. An illustration of proposed network architecture is in Fig. 2.

3.1 DenseNet for Chest X-rays

Fully convolutional network (FCN) [24] has become one of the most successful deep learning frameworks for generic image classification and segmentation tasks. In [33], ResNet, a recent FCN alike model, delivered best classification accuracy on ChestX-ray8. A typical DenseNet [14] comprises multiple densely connected convolutional layers, which improve the flow of information and gradients through the network, making it converge better and mitigating gradient vanishing issue. Therefore, in many computer vision tasks, DenseNet has shown magnificently stronger capability of representation learning than ResNet. Then [25] fine-tuned a DenseNet that naturally preserves spatial information throughout the network. As well as on the purpose of a fair comparison, we particularly choose the publicly available DenseNet-121 model as backbone network [1]. As shown in Fig. 2, the backbone of the used DenseNet consists of four consecutive dense blocks. However, original DenseNet is incapable of handling the special issues in disease classification and localization on chest X-rays. For example, disease labels of ChestX-ray14 are highly noisy since they were generated from scanning report. Given a X-ray image corresponds to multiple disease types, it is still an open question how to make data selectively contribute to multiple classification and localization tasks.

3.2 Squeeze-and-Excitation Block in DenseNet

In classical CNNs, it is difficult to model the interdependency between channels using convolutional filters, which are initialized and

[1] https://github.com/pytorch/vision/blob/master/torchvision/models/densenet.py

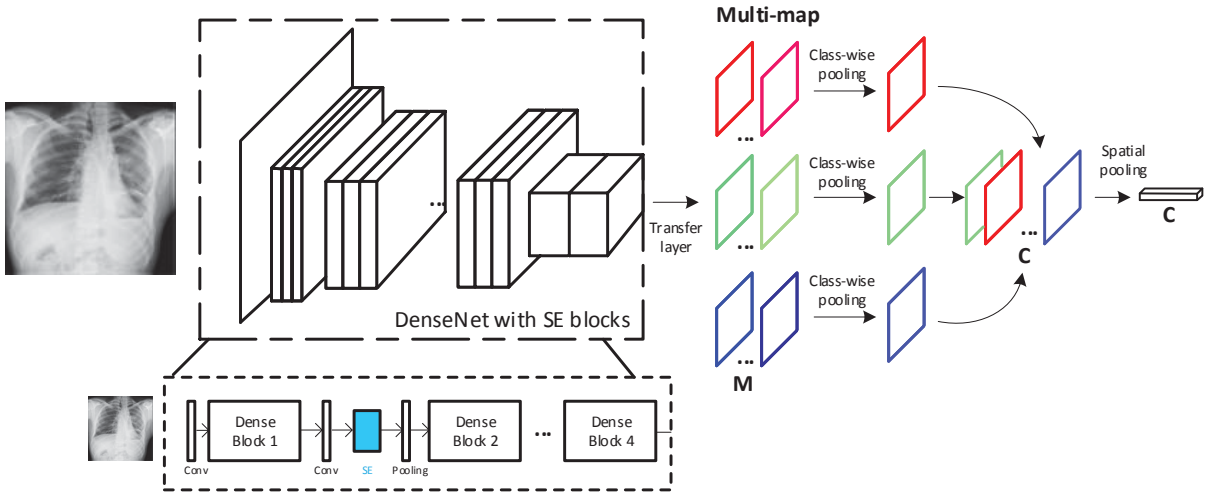

Figure 2: The Proposed Network Architecture.

trained independently. However, the cross-channel dependency is widely existing and has been recognized as one of the major visual patterns, e.g. joint sparsity [21].

In between two consecutive dense blocks of DenseNet, there is a convolution-pooling operator that transforms previous activation output to a new feature space and then squeezes it to a compact spatial domain. In proposed model, we insert a so-called *squeeze-and-excitation* (SE) block into the convolution-pooling operator. Particularly, we first squeeze the C feature maps after convolution into a feature vector of C length by spatial average-pooling. An *excitation* process is to reweight feature maps by the channel-wise attention coefficients learned from the squeezed vector. The motivation is to offer a chance of cross-channel feature recalibration considering the channel interdependencies.

Squeeze Before recalibration, we need a global statistic of each channel. Then a global squeezing is performed first by an average-pooling across entire spatial domain. Consider $\mathbf{U} \in \mathbf{R}^{H\times W\times C}$ as transformed feature maps after convolution, where $H\times W\times C$ is the dimensionality. A *squeeze* operation is to aggregate the feature maps across spatial dimensions $H \times W$ to produce a channel descriptor forming a C-length descriptor vector for entire $\mathbf{U}$. Assume $\mathbf{z}$ is the vector after squeezing and the c-th element of $\mathbf{z}$ is calculated by

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

This was not possible in classical CNN in which feature maps were convolved independently by separate filter kernels and therefore the squeezing scale was constrained within reception field and the pooling was also committed locally.

Excitation To recalibrate feature maps channel-wise, we need to learn the channel weights. We employ a self-gating mechanism, which outputs channel attentions, based on the non-linear channel interdependence after passing a *sigmoid* activation function σ:

$$\mathbf{s} = \sigma(\mathbf{W}_2 \times ReLU(\mathbf{W}_1 \times \mathbf{z}))\,, \tag{2}$$

where $\mathbf{s} \in \mathbf{R}^C$ is the channel-wise attention coefficients for feature recalibration. Due to Eq 2, channel coefficient s^c represented the relative importance of channel c. For the purpose of reducing complexity, a bottleneck structure formed by two fully connected layers parameterised by $\mathbf{W}_1 \in \mathbf{R}^{\frac{C}{r}\times C}$ and $\mathbf{W}_2 \in \mathbf{R}^{C\times\frac{C}{r}}$ (r is the reduction ratio) is used in Eq 2 to adaptively adjust channel importance according to learning objective. The final output after SE block of channel c, $\tilde{x_c}$, is obtained by re-scaling the transformed feature maps $\mathbf{U}$ with $\mathbf{s}$ by a channel-wise multiplication:

$$\tilde{x_c} = s^c \cdot u^c, \quad c \in \{0, \ldots, C-1\}. \tag{3}$$

The physical meaning of SE block for classification of chest X-rays comes from the hardly distinguishable illuminative contrast between lesion regions of different types of disease as well as the rest normal regions. Therefore, merely utilizing single feature map or independently processing multiple maps cannot provide enough informative features for disease classification. The workflow of SE block is given in Fig.3.

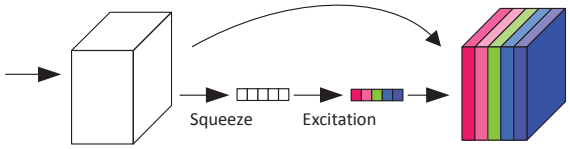

Figure 3: Illustration of a Squeeze-and-Excitation Block.

3.3 Multi-map Layer and Max-min Pooling

Because ChestX-ray14 offers multiple disease labels for most of X-rays, it is naturally required to perform a multi-class classification.

Instead of generating a multi-hot score vector, which makes training difficult to converge, we were encouraged by good performance from introducing multi-map transfer layer, each output feature map of which corresponds to a particular disease class.

The last dense block generates feature maps with size as $w \times h \times d$. Then we concatenate to it a multi-map transfer layer. The layer encodes the activation outputs of backbone network into M individual feature maps for each disease class through 1×1 convolution operation. Denote M as the number of feature maps per class and C as the number of classes, this transfer layer will achieve the output of size $w \times h \times MC$. When $M = 1$, it is reduced to a standard classification output of C classes. The modalities are learned with only image-level label and the transfer layer maintains spatial resolution. The M modalities aim at specializing to different class-related visual features.

To sufficiently utilize the provided multi-class label, we proposed a two-stage pooling layer to aggregate information on feature maps for each disease class. A standard class-wise average-pooling was first conducted to transform maps from $w \times h \times MC$ to $w \times h \times C$. As to spatial aggregation, we applied a recently proposed spatial max-min pooling [7] to globally extract spatial domain information. Because we find that global minimum information is also helpful for the medical image analysis, and the global minimum regions can act as a regularizer and reduce overfitting. The global maximum and minimum pooling are linearly combined in our model:

$$r^c = \max_{\mathbf{h} \in \mathcal{H}_{k^+}} \frac{1}{k^+} \sum_{i,j} h_{i,j} \bar{z}^c_{i,j} + \alpha \left(\min_{\mathbf{h} \in \mathcal{H}_{k^-}} \frac{1}{k^-} \sum_{i,j} h_{i,j} \bar{z}^c_{i,j} \right), \quad (4)$$

where $\bar{z}^c$ is the c-th pooled feature map after class-wise pooling. $\mathcal{H}_k$ is the set that $\mathbf{h} \in \mathcal{H}_k$ satisfies $h_{i,j} \in \{0, 1\}$ and $\sum_{i,j} h_{i,j} = k$. The max-min spatial pooling consists in selecting for each class the positive k^+ regions with the highest activations from input $\bar{z}^c$ and vice versa. The output r^c is the weighted average of scores of all the selected regions. To generate the final positive probability, we pass r^c through a sigmoid activation function.

3.4 Comparison with CheXNet

The CheXNet [25] is a similar model that also uses DenseNet-121 as the backbone network. It removes the last linear layer of DenseNet and adds a 1×1 convolutional layer as the transfer layer to convert the extracted 1024-channel feature maps into C-channel feature maps. To get the final C-dimensional output, it then uses the global maximum pooling and the sigmoid function. Compare with the CheXNet, the proposed architecture completely remove the liner layer and is fully convolutional. Our model have significant modifications particularly for learning representations under a weak supervision. We highlight the significant modifications as below:

- Make the model fully convolutional by removing the linear layer. Fully convolutional architecture is suitable for spatial learning.
- SE blocks perform feature recalibration by weights learned from channel interdependencies, improving the representational power of CheXNet.
- Different from CheXNet that still only has one single feature map, we use multi-map transfer layer to encode modalities associated with each individual disease class, making our framework more capable of discriminating the appearance of multiple thoracic diseases on the same chest X-ray.
- To aggregate spatial scores from multi-maps into a global prediction, We incorporate a novel max-min pooling strategy which is better than the global pooling in CheXNet.

4 EXPERIMENT

4.1 Chest X-ray Dataset

The problem of thoracic disease classification and detection on chest X-rays has been extensively explored. Recently, Wang et al. [33] released the largest chest X-ray dataset so far - ChestX-ray14, which collects 112,120 frontal-view chest X-ray images of 30,805 unique patients. Each radiography is labeled with one or multiple types of 14 common thorax diseases: Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, Pneumothorax, Consolidation, Edema, Emphysema, Fibrosis, Pleural Thickening and Hernia. These disease labels were mined from the associated radiological reports (> 90 % accuracy [33]). Besides, there are 880 X-rays provided with lesion regions annotated as bounding boxes by radiologists. In our experiments, we only used disease label as ground-truth in training and evaluating the model in disease classification. We also utilized the bounding boxes only for a visual evaluation of disease region localization on X-rays.

To have a fair comparison with previous methods [25, 33, 34], we splitted the dataset into three parts: training, validation, and evaluation, on patient level using the publicly available data split list [33]. There are respectively 76,524, 10,000, and 25,596 chest X-ray images for training, validation, and evaluation purposes. Since there may be multiple X-rays for each patient, split on patient level can guarantee the X-rays of the same patient be assigned to the same part. Split on image level will introduce potential over-fitting since the X-rays of the same patient can be assigned to both training and evaluation subsets.

4.2 Experimental Setting

Similar to [25, 33, 34], we formulate the Chest X-ray disease recognition as a classical multi-label classification problem. The proposed model outputs a 14-dimensional vector indicating the positive probability for each kind of listed diseases. An all-zero vector represents normal status (None of 14 listed thoracic diseases are detected). We use the standard binary cross entropy loss as objective function. ROC- AUC score (the area under the Receiver Operating Characteristic curve) are used as evaluation metric in disease classification.

For SE blocks, we set the reduction ratio to be 16 as suggested in [13]. We set M in the multi-map layer as 12, which was experimentally proved to be an effective trade-off between the performance and the complexity. For max-min pooling, we use $k^+ = k^- = 1$ and $\alpha = 0.7$ as given in [7]. The end-to-end model was trained by Adam optimizer [17] with standard parameters ($\beta_1 = 0.9$ and $\beta_2 = 0.99$). We initialize the model using weights from the pre-trained DenseNet model, and only train the multi-map transfer layer and newly inserted Squeeze-and-Excitation layer from scratch. Following a previous work on ChestX-ray14 [25], we set the batch size 16 and initial learning rate 0.0001. The learning rate will be decayed by 10 times when the validation loss plateaus for more than 5 epochs. The model of the least validation loss will be the selected classifier.

Table 1: The comparison of AUC scores. The best AUC score in each row is displayed in bold. Note that Li et al.[22] used extra disease location information when training the model and did not perform on official split.

	ChestX-ray8 [33]	Yao et al. [34]	Li et al. [22]	DNetLoc [10]	CheXNet [25]	Our Method
Official Split	Yes	Yes	No	Yes	Yes	Yes
Atelectasis	0.7160	0.7330	**0.8000**	0.7670	0.7795	0.7924
Cardiomegaly	0.8070	0.8580	0.8700	**0.8830**	0.8816	0.8814
Effusion	0.7840	0.8060	**0.8700**	0.8280	0.8268	0.8415
Infiltration	0.6090	0.6750	0.7000	0.7090	0.6894	**0.7095**
Mass	0.7060	0.7270	0.8300	0.8210	0.8307	**0.8470**
Nodule	0.6710	0.7780	0.7500	0.7580	0.7814	**0.8105**
Pneumonia	0.6330	0.6900	0.6700	0.7310	0.7354	**0.7397**
Pneumothorax	0.8060	0.8050	0.8700	0.8460	0.8513	**0.8759**
Consolidation	0.7080	0.7170	**0.8000**	0.7450	0.7542	0.7598
Edema	0.8350	0.8060	**0.8800**	0.8350	0.8496	0.8478
Emphysema	0.8150	0.8420	0.9100	0.8950	0.9249	**0.9422**
Fibrosis	0.7690	0.7570	0.7800	0.8180	0.8219	**0.8326**
Pleural Thickenin	0.708	0.7240	0.7900	0.7610	0.7925	**0.8083**
Hernia	0.7670	0.8240	0.7700	0.8960	0.9323	**0.9341**
Average	0.7381	0.7673	0.8064	0.8066	0.8180	**0.8302**

The original image size 1024×1024 is infeasible for a very deep convolutional neural network. In this paper, we resize the images to be of size 512×512 and convert single channel X-ray images into 3-channel RGB images since the pre-trained DenseNet only accepts 3-channel images as input. As ImageNet [6] the pixel values in each channel were normalized. During training, we randomly crop a 448×448 sub-image from the input 512×512 image to augment the original training subset. The cropped sub-image is randomly horizontally flipped to incrementally increase the variation and the diversity of training samples. During the evaluation process, we use as input ten randomly cropped 448×448 sub-images (four corner crops and one central crop plus horizontally flipped version of these) for each evaluation sample, and take the average probability as the final prediction.

4.3 Comparison with State-of-the-art Methods

We compared the classification performance of our proposed model with previously published methods, including Wang [33], Li Yao [34], DNetLoc [10], ChexNet [25] and Zhe Li [22]. We showed that our method achieved current state-of-the-art classification accuracy on ChestX-ray14 dataset. In the experiments, we found that different data split setup has significant influence on the model performance. However, the results of ChexNet in [25] was not achieved under the official data splitting. To make a fair comparison, we implement the ChexNet and evaluate its performance with the provided official data split. It is noted that Li et al. [22] used extra disease location information than others in training and did not use the official split. Therefore, it is not comparable to our method as well as other state-of-the-arts approaches. Even though, we still outperformed [22] in classification of 9 out of the 14 diseases.

Numerical classification results are given in Table 1. For each evaluated method, we report ROC-AUC scores for each disease class as well as the average score of all classes. Compared with previous methods, our network improves the overall performance by 2%. Especially, for some challenging diseases, e.g. "Lung Nodule", the accuracy was dramatically improved by a margin of at least 3%. The performance is generally improved because of the better spatial squeezing capability from the use of SE blocks and the max-min pooling operation. Moreover, for the same reason, our method can effectively handle lesion areas of different size. For example, "Cardiomegaly" and "Edema" have relatively larger pathology areas on X-rays than "Mass" and "Nodule". From Table 1, it is verified that the proposed network can effectively learn decisive features from X-rays of both large and small disease areas, while others cannot.

4.4 Localization of Lesion Regions

In Fig.4, we produce heat map to visualize the most indicative pathology areas on X-rays from evaluation subset, interpreting the representational power of network. Heat maps are constructed by computing the average of class-wise features after pooling along the channel dimension [9]. We can see that our proposed network is able to localize lesion region on X-rays by assigning higher values than the normal. A visual evaluation has confirmed that the highlighted regions on X-rays are pretty close to ground-truth (red bounding boxes). Since our model did not use any bounding boxes in training, this has demonstrated that the proposed framework has a good interpretation ability in terms of localizing disease regions and can be widely applied in clinical practice where detailed annotations are hardly available.

4.5 Ablation Study

In the section, we conduct additional ablation experiments to demonstrate the effectiveness of three proposed components in our network that respectively bring performance gains: multi-map transfer layer, max-min pooling and SE block. From Table.1, CheXNet has the average AUC score as 0.8180 for all 14 diseases. The average AUC score of our method is 0.8302. This 1.2% lift demonstrated the

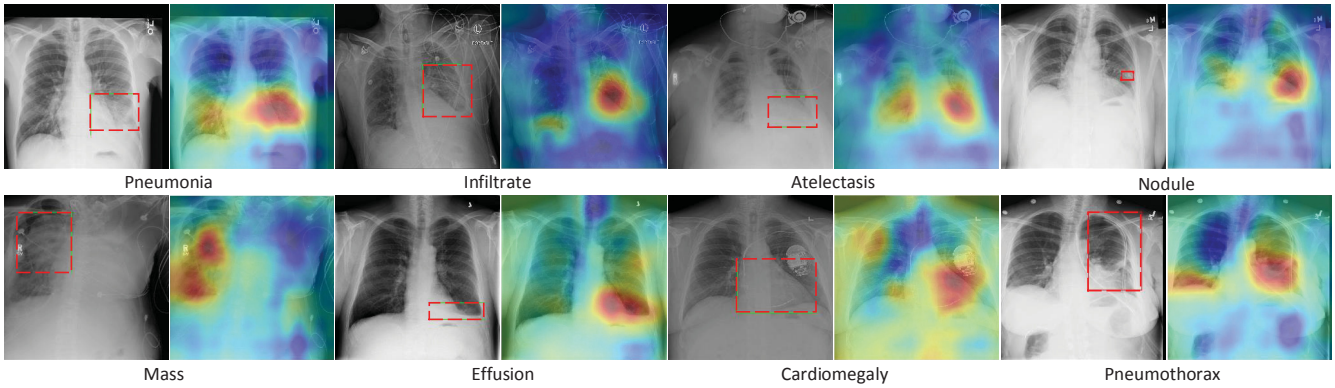

Figure 4: The proposed method localizes the areas of the X-ray that are most important for making particular pathology classification. We can see that the localized areas are very close to the corresponding bounding boxes.

Table 2: Validation of the effectiveness of the three improvements. The best AUC score in each row is displayed in bold.

	Our Method	w/o SE	w/o multi-map	w/o max-min pooling
Atelectasis	**0.7924**	0.7867	0.7900	0.7784
Cardiomegaly	0.8814	**0.8852**	0.8790	0.8762
Effusion	0.8415	0.8418	**0.8420**	0.8392
Infiltration	**0.7095**	0.7048	0.7087	0.6985
Mass	**0.8470**	0.8462	0.8469	0.8440
Nodule	0.8105	0.8055	**0.8110**	0.8034
Pneumonia	**0.7397**	0.7368	0.7364	0.7435
Pneumothorax	**0.8759**	0.8738	0.8736	0.8753
Consolidation	0.7598	**0.7640**	0.7586	0.7545
Edema	0.8478	0.8464	**0.8503**	0.8398
Emphysema	0.9422	0.9402	**0.9436**	0.9371
Fibrosis	**0.8326**	0.8269	0.8302	0.8067
Pleural Thickenin	0.7994	**0.8059**	0.8058	0.8011
Hernia	**0.9341**	0.9330	0.9299	0.9096
Average	**0.8302**	0.8279	0.8290	0.8220

joint effects from the three components compared to the state-of-the-art. Now we will validate the difference on performance of our model when sequentially removing each contributive component. It is noted that, in the next three experiments, we only changed the network structure and keep other experimental setting identical to make a fair and illustrative comparison. Results are in Table.2.

Effectiveness of SE Block In the original DenseNet architecture, the transition layer between consecutive dense blocks is simply a 1×1 convolutional layer followed by a average-pooling layer for purpose of dimension reduction. In our model, we realized that *squeeze* operation will extend spatial aggregation to the entire spatial domain, which was impossible for a local pooling. Besides, *excitation* process will train a parameterised reweighting on feature maps supervised by channel interdependencies, which was also not possible in previous transition layer of DenseNet. When we remove SE blocks from the network, the average AUC score drops to 0.8279 showing that our method indeed achieves performance gain by using SE blocks as they recalibrate convolutional features.

Effectiveness of Multi-map Transfer Layer We use multiple feature maps for each disease in our model. In experiments, the learning of multi-map was skipped by setting $M = 1$. Consequently, the average score of revised model becomes 0.8290. As shown in Fig. 2, the appearances of different classes of disease vary a lot in shape and color, which supported the use of multi-map transfer layer.

Effectiveness of Max-min Pooling CheXNet adopted traditional global maximum pooling which only extracts the maximum component for the whole feature map assuming the maximum component is considered to be the most informative part. However, we found that the minimum components also contribute a lot to the thoracic disease classification. Results from the Table.2 validated the effectiveness of max-min pooling showing that our model would lose 1.0% on ROC-AUC score when using only maximum pooling.

5 CONCLUSION

In this paper, we proposed a unified weakly-supervised deep learning framework to jointly perform thoracic disease classification and localization on chest X-rays only using noisy multi-class disease label. The advantages of proposed network are not only from the learning of disease-specific features via multi-map transfer layers, also from the cross-channel feature recalibration by *sqeeuze-and-excitation* blocks in between dense blocks. Heat maps, as by-product obtained under weak supervision, further visualize the representational power of our network. This also highlights the interpretability of our model. Finally, both quantitative and qualitative results has indicated that our framework outperformed the state-of-the-arts. As to future work, we will re-investigate an accurate localization of lesion areas utilizing the limited amount of bounding boxes.

REFERENCES

[1] Alan R Aronson and François-Michel Lang. 2010. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association* 17, 3 (2010), 229–236.

[2] Wendy Webber Chapman, Marcelo Fizman, Brian E Chapman, and Peter J Haug. 2001. A comparison of classification algorithms to automatically identify chest X-ray reports that support pneumonia. *Journal of biomedical informatics* 34, 1 (2001), 4–14.

[3] C. Chen, V. C. Kavuri, X. Wang, R. Li, H. Liu, and J. Huang. 2015. Multi-frequency diffuse optical tomography for cancer detection. In *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*. 67–70. https://doi.org/10.1109/ISBI.2015.7163818

[4] David J Crandall and Daniel P Huttenlocher. 2006. Weakly supervised learning of part-based spatial models for visual object recognition. In *European conference on computer vision*. Springer, 16–29.

[5] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. 2015. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association* 23, 2 (2015), 304–310.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 248–255.

[7] Thibaut Durand, Taylor Mordan, Nicolas Thome, and Matthieu Cord. 2017. Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*.

[8] Yu Gordienko, Peng Gang, Jiang Hui, Wei Zeng, Yu Kochura, O Alienin, O Rokovyi, and S Stirenko. 2017. Deep Learning with Lung Segmentation and Bone Shadow Exclusion Techniques for Chest X-Ray Analysis of Lung Cancer. *arXiv preprint arXiv:1712.07632* (2017).

[9] Qingji Guan, Yaping Huang, Zhun Zhong, Zhedong Zheng, Liang Zheng, and Yi Yang. 2018. Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification. *arXiv preprint arXiv:1801.09927* (2018).

[10] Sebastian Guendel, Sasa Grbic, Bogdan Georgescu, Kevin Zhou, Ludwig Ritschl, Andreas Meier, and Dorin Comaniciu. 2018. Learning to recognize abnormalities in chest x-rays with location-aware dense networks. *arXiv preprint arXiv:1803.04565* (2018).

[11] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 316, 22 (2016), 2402–2410.

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

[13] Jie Hu, Li Shen, and Gang Sun. 2017. Squeeze-and-excitation networks. *arXiv preprint arXiv:1709.01507* (2017).

[14] Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, Vol. 1. 3.

[15] Mohammad Tariqul Islam, Md Abdul Aowal, Ahmed Tahseen Minhaz, and Khalid Ashraf. 2017. Abnormality Detection and Localization in Chest X-Rays using Deep Convolutional Neural Networks. *arXiv preprint arXiv:1705.09850* (2017).

[16] Simon Jégou, Michal Drozdzal, David Vazquez, Adriana Romero, and Yoshua Bengio. 2017. The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*. IEEE, 1175–1183.

[17] Diederik P Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *Computer Science* (2014).

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

[19] Robert Leaman, Ritu Khare, and Zhiyong Lu. 2015. Challenges in clinical natural language processing for automated disorder normalization. *Journal of biomedical informatics* 57 (2015), 28–37.

[20] Ruoyu Li and Junzhou Huang. 2015. Fast regions-of-interest detection in whole slide histopathology images. In *International Workshop on Patch-based Techniques in Medical Imaging*. Springer, 120–127.

[21] Ruoyu Li, Yeqing Li, Ruogu Fang, Shaoting Zhang, Hao Pan, and Junzhou Huang. 2015. Fast preconditioning for accelerated multi-contrast MRI reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 700–707.

[22] Zhe Li, Chong Wang, Mei Han, Yuan Xue, Wei Wei, Li-Jia Li, and Fei-Fei Li. 2017. Thoracic Disease Identification and Localization with Limited Supervision. *arXiv preprint arXiv:1711.06373* (2017).

[23] Yun Liu, Krishna Gadepalli, Mohammad Norouzi, George E Dahl, Timo Kohlberger, Aleksey Boyko, Subhashini Venugopalan, Aleksei Timofeev, Philip Q Nelson, Greg S Corrado, et al. 2017. Detecting cancer metastases on gigapixel pathology images. *arXiv preprint arXiv:1703.02442* (2017).

[24] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3431–3440.

[25] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. 2017. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. *arXiv preprint arXiv:1711.05225* (2017).

[26] Jonathan Rubin, Deepan Sanghavi, Claire Zhao, Kathy Lee, Ashequl Qadir, and Minnan Xu-Wilson. 2018. Large Scale Automated Reading of Frontal and Lateral Chest X-Rays using Dual Convolutional Neural Networks. *arXiv preprint arXiv:1804.07839* (2018).

[27] Arnaud Arindra Adiyoso Setio, Alberto Traverso, Thomas De Bel, Moira SN Berens, Cas van den Bogaard, Piergiorgio Cerello, Hao Chen, Qi Dou, Maria Evelina Fantacci, Bram Geurts, et al. 2017. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge. *Medical image analysis* 42 (2017), 1–13.

[28] Junji Shiraishi, Shigehiko Katsuragawa, Junpei Ikezoe, Tsuneo Matsumoto, Takeshi Kobayashi, Ken-ichi Komatsu, Mitate Matsui, Hiroshi Fujita, Yoshie Kodera, and Kunio Doi. 2000. Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *American Journal of Roentgenology* 174, 1 (2000), 71–74.

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

[30] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, et al. 2015. Going deeper with convolutions. Cvpr.

[31] Lorenzo Torresani. 2014. Weakly supervised learning. In *Computer Vision*. Springer, 883–885.

[32] Bram Van Ginneken, Mikkel B Stegmann, and Marco Loog. 2006. Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database. *Medical image analysis* 10, 1 (2006), 19–40.

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

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

[35] Li Yao, Jordan Prosky, Eric Poblenz, Ben Covington, and Kevin Lyman. 2018. Weakly Supervised Medical Diagnosis and Localization from Multiple Resolutions. *arXiv preprint arXiv:1803.07703* (2018).

[36] Xujiong Ye, Xinyu Lin, Jamshid Dehmeshki, Greg Slabaugh, and Gareth Beddoe. 2009. Shape-based computer-aided detection of lung nodules in thoracic CT images. *IEEE Transactions on Biomedical Engineering* 56, 7 (2009), 1810–1820.

[37] Xinliang Zhu, Jiawen Yao, Feiyun Zhu, and Junzhou Huang. 2017. Wsisa: Making survival prediction from whole slide histopathological images. In *IEEE Conference on Computer Vision and Pattern Recognition*. 7234–7242.