Weakly supervised classification and localization of common thorax diseases

Shivam Mittal
2015csb1032@iitrpr.ac.in
Department of Computer Science,
IIT Ropar

Abhinav Dhall
abhinav@iitrpr.ac.in
Department of Computer Science,
IIT Ropar

1 Abstract

In this project, we use the NIH Chest X-ray Dataset which comprises of over 1 lakh frontal-view X-ray images with labels for each image specifying the presence/absence of 14 common thorax diseases. We develop a model for multi label image classification and weakly supervised pathology localization (weakly supervised because there are no bounding box labels for training images), which detects the presence of multiple diseases and also generates bounding boxes around the corresponding disease. Classifying and localizing the diseases using weak supervision within the x-ray is an challenging task, with many practical application including increasing the accessibility of better health care service in remote areas. The model we develop is a CNN (VGG-net) for handling the multi-label classification problem. Using a simple modification of adding a global average pooling layer, we can use the feature maps and weights of the network to generate a likelihood map of where the pathology can be located. Using this heatmap, bounding boxes localizing the pathology are generated for each x-ray image.

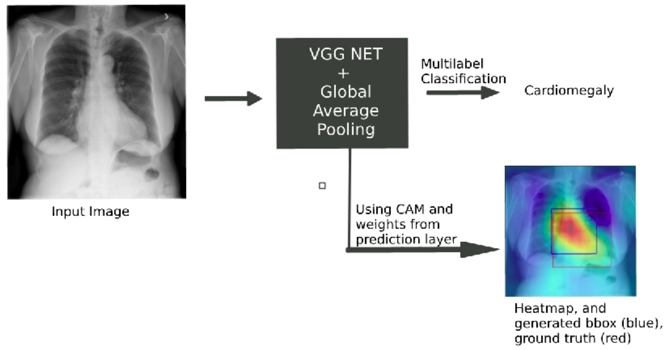

Figure 1: Given an input x-ray image (containing pathology cardiomegaly), our model produces the multilabel classification of the pathology (more than one can be present) present in the xray, and the bounding box localizing the pathology based on the generated heat map.

2 Introduction

A chest x-ray is one of the most common non invasive radiology test which produces an image of the chest and thae internal organs. These x-ray images are black and white with only the brightness defining the structure of organs. This radiology test contains a lot of information and is commonly used to detect clinical conditions such as pneumonia, mass, nodules, etc.

Computer Aided Diagnosis (CAD) has been a sought after research field. Automating the task of detecting diseases within a chest x-ray can be of great use practically. It can be used as a assistive technology for radiologists to speed up their work, prevent errors and increase productivity. Also, it can be used to increase the accessibility of better health care service in remote areas.

However classifying and localizing the diseases within the x-ray is an challenging task because the disease (or the object which has to be detected) is very small in size compared to the whole x-ray image. Also, there is no fully labelled large dataset available with bounding box information of the diseases, so framing this problem as supervised task is not possible. Wang et al [21] assembled a dataset named Chest X-ray dataset which consists over 1 lakh frontal view chest x-ray images along with their text-mined disease labels. The dataset also contains a small subset of images with region-level annotations (bounding boxes) for validation and testing purposes.

Given the chest x-ray dataset, the problem is formulated as a weakly supervised multi-label classification and localization problem. The task is given an input x-ray image, we have to detect the diseases (multi-label) present and also localize (draw a bounding box around) around the diseases. It is a weakly supervised task because in the training set, only class labels are present for each image, the bounding box information is not present.

Recent years have shown the use of deep learning algorithms being applied in the field of medical image analysis. Examples include predicting spinal radiological scores, segmenting cartilage from human knee, pancreas segmentation, etc.

But still building stand alone medical imaging systems is a very challenging task because predicting false negatives and false positives have serious consequences attached with them.

This task is more challenging than traditional computer vision tasks because usually anatomical structures do not have very well defined edges when captured from most imaging modalities. This makes problems such as segmentation or localization harder.

However, the performance of such systems are increasing as more datasets are being publically available, and better deep learning models are being built.

The challenge for this task is two fold. First is to be able to design a model powerful enough to learn the features necessary for detecting and describing the diseases which may be small in size compared to the whole image. The Convolution Neural Networks (CNN) find application in this prospect. CNN have evolved with time to best describe the images, specifically in the task of image classification, with each layer depicting some information regarding the image, as we go deeper into the layered structure, information at each level starts holding much complex and useful relations and meaning with respect to the image. Hence the deeper layers of CNN becomes an obvious and inevitable source to depict the characteristics of the image. The second task is to localize the diseases with weak supervision (having no region level annotations in the training data). There has been work [6] [16] [23] [15] [2] in this area which will be discussed in further sections which shows that this task can be done with reasonable accuracy.

We develop a CNN model using pre-trained VGG-net (pre-trained on imagenet), and making slight modification to the network by adding global average pooling layer, which allows us to use the feature maps and the weights of the fine-tuned network (trained on the x-ray data) to generate heatmaps which specify the likelihood of a disease being present. We'll describe the model in detail in further sections.

3 Related work

Weakly supervised classification and localization is a fairly well studied problem in recent times. Some recent attempts [12] [14] [18] have also been made to classifying the dataset which we would be dealing with.

Initial work [5] [1] [10] [7] [22] on weakly supervised object localization has focused on learning from images containing prominent and centered objects in scenes with limited background clutter. More recent efforts attempt to learn from images containing multiple objects embedded in complex scenes [20] [17] [8] [4] [6]. These methods typically aim to localize objects including finding their extent in the form of bounding boxes. They attempt to find parts of images with visually consistent appearance in the training data that often contains multiple objects in different spatial configurations and cluttered backgrounds. While these works are promising, their performance is still far from the fully supervised methods such as [11] [19].

We would be describing the most recent work on weakly supervised object localization. Bazzani et al [2] introduce a technique self-taught object localization that leverages deep convolutional networks trained for whole-image recognition to localize objects in images without additional human supervision, i.e., without using any ground-truth bounding boxes for training. The key idea that they use is to analyze the change in the recognition scores when artificially masking out different regions of the image. The masking out of a region that includes the object typically causes a significant drop in recognition score. This idea is embedded into an agglomerative clustering technique that generates self-taught localization hypotheses. Cinbis et al [6] follow a multiple-instance learning approach that iteratively trains the detector and infers the object locations in the positive training images. Oquab et al [15] propose a method for transferring mid-level image representations and show that some object localization can be achieved by evaluating the output of CNNs on multiple overlapping patches. However, the authors do not actually evaluate the localization ability. On the other hand, while these approaches yield promising results, they are not trained end-to-end and require multiple forward passes of a network to localize objects, making them difficult to scale to real-world datasets. Oquab et al [16] use the successful CNN architecture which shows state-of-the-art results for object classification, and modify the network such as treating the last fully connected layer as convolutions to remove the uncertainty in object localization, they also introduce a global max-pooling layer that hypothesizes the possible location of the object in the image. This network when trained to output image-level labels only, also localizes objects or their distinctive parts in training images. Zhou et al [23] use a similar approach to Oquab et al. They remove the fully connected layer, and train a CNN with a global average pooling layer at the end. Then using the activation maps and the class prediction score, class activation maps can be obtained by just training the CNN on object categorization. These class activation maps are like a heat map indicating the locations where the object is most likely present.

Also, people have been leveraging the power of GANs to improve upon localization techniques. Some of the notable (not all included) works are described below. Behpour et al [3] propose deep adversarial object localization, which approximates ground truth annotations of training images instead of approximating the loss function by posing object localization as an adversarial game between a loss-minimizing prediction player and a loss-maximizing adversarial player. They constrain the adversary to match specified properties of training data that are uncovered from a convolutional neural networkâĂŹs feature representation. Diba et al [9] approach GANs with a novel training method and learning objective, to discover multiple object instances for three cases: 1) synthesizing a picture of a specific object within a cluttered scene; 2) localizing different categories in images for weakly supervised object detection; and 3) improving object discovery in object detection pipelines. A crucial advantage of their method is that it learns a new deep similarity metric, to distinguish multiple objects in one image. Li et al [13] propose a new Perceptual Generative Adversarial Network (Perceptual GAN) model that improves small object detection through narrowing representation difference of small objects from the large ones. Specifically, its generator learns to transfer perceived poor representations of the small objects to super-resolved ones that are similar enough to real large objects to fool a competing discriminator. Meanwhile its discriminator competes with the generator to identify the generated representation and imposes an additional perceptual requirement generated representations of small objects must be beneficial for detection purpose on the generator.

There has been work which classified the dataset that we will be using. Wang et al [21] were the ones who had assembled the dataset. They presented a unified weakly-supervised multi-label image classification and pathology localization framework, which can detect the presence of multiple pathologies and subsequently generate bounding boxes around the corresponding pathologies. They train a Deep Convolution Neural Network on the disease classification task, and then use the feature maps and the weights of prediction layer to find the plausible spatial location of diseases by generating a likelihood map of pathologies as described in [23]. [12] and [18] use deep learning models on the dataset but only for the classification task, they do not perform the localization task. Li et al [14] crosses the baseline set by Wang et al. They first apply a CNN to the input image so that the model learns the information of the entire image and implicitly encodes both the class and location information for the disease. Then they slice the image into a patch grid to capture the local information of the disease. The task is then formulated as a multiple instance learning (MIL) problem, at least one patch in the image belongs to that disease. If there is no disease in the image, all patches have to be disease free. In this way, they have unified the disease identification and localization into the same underlying prediction model.

4 Model

We'll describe the unified model we developed for weakly supervised multi-label classification and pathology localization task. The training and the inference framework are described separately in figure 2 and figure 3 respectively.

4.1 The CNN framework

The convolutional model which we used is described in figure 2 and figure 3. The initial layers of the network were taken from pre-trained VGG-network pre-trained on Imagenet data. We remove the fully connected layers present in VGG-net so that the spatial location information is maintained in the feature maps. After the VGG convolutional and max-pool layers, we add another convolutional layer to increase the depth of the feature maps. After that we add a global average pooling layer, a prediction layer (fully connected layer) and a loss layer at the end.

The whole network is used for the multi-label classification task, and subsequently we can use the features obtained from the forward pass and weights of the network to obtain the heatmap as shown in figure 3.

4.1.1 Multi-label setup

For representing the labels of the images, we define a 14 dimensional vector $y = [y_1, y_2, ..., y_{14}]$ where each $y_i \in \{0,1\}(1 <= i <= 14)$. Each y_i represents the presence/absence of one disease. The presence of no disease or "Normal" is represented by an all zero vector.

4.1.2 Loss function

We use the sigmoid cross entropy loss, i.e. apply sigmoid function on the scores obtained from the network to scale them between 0 and 1, and then apply cross entropy loss between the obtained probabilities and the true label.

$Loss = sum(z * -log(sigmoid(x)) + (1-z) * -log(1-sigmoid(x)))$

where x=logits (scores from network) and z = labels.

4.2 Localization task

The main framework is described in figure 3. The individual parts will be described in greater details in the following sections.

4.2.1 Heatmap/CAM generation

A class activation mapping or the heatmap for a particular class basically identifies those discriminative spatial locations which are used by the CNN to identify the that class.

The initial parts of the network consists of convolutional and max-pool layers to give convolutional feature map which is fed to a global average pooling layer. Global average pooling averages across one 2-D feature map to generate one value. So a $S*S*D$ where D is the number of channels in the feature map gets reduced to $1*D$. This $1*D$ vector is used as a feature for a fully connected layer which produce the final classification scores. The classification scores are produced by the weighted sum (through the weights of the fully connected layer) of the $1*D$ vector. Similarly using the same weights, the weighted sum of the feature maps of the last convolutional layer are used to generate the class activation maps (CAM).

This process is also explained in figure 3.

4.2.2 Bounding box generation

The heatmaps produced from the framework indicate the spatial locations which are likely to contain the pathology. We normalize the scores of the heatmap between 0 and 1, and apply a simple threshold to consider the values above 0.65 (determined empirically) only. Then, we generate a

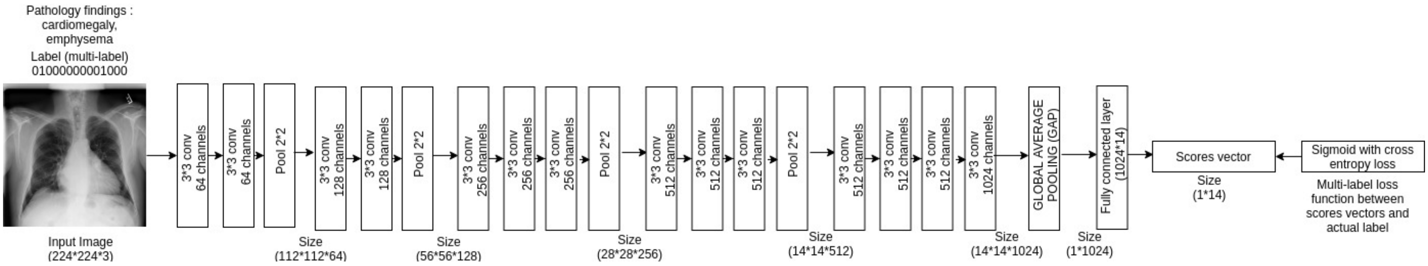

Figure 2: Our model, and the training process

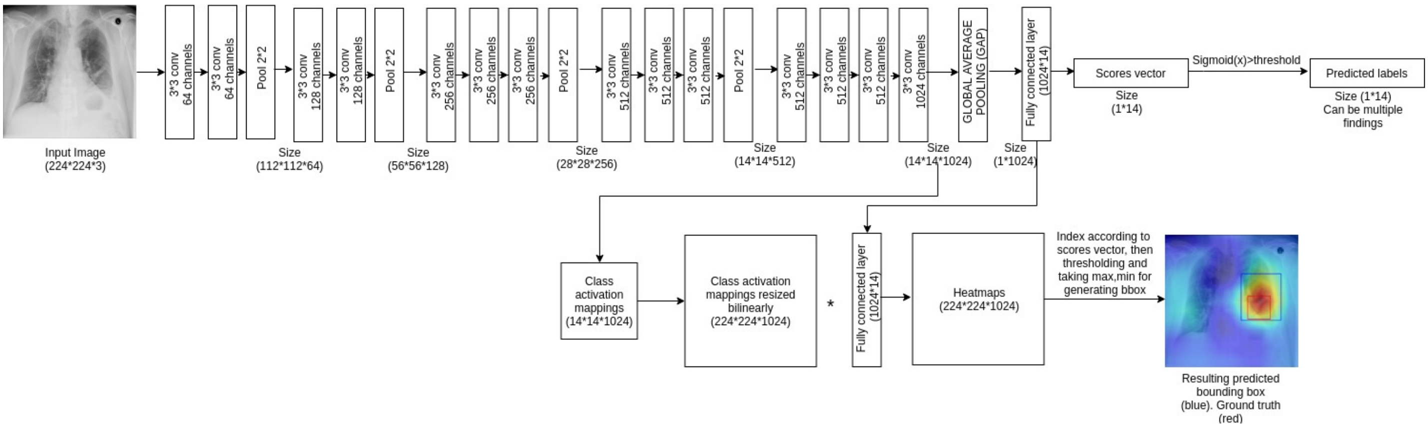

Figure 3: Our model, and the inference process

bounding box which covers all those locations in the heatmap which had a score above the threshold (0.65).

5 Experiments and results

We'll define the dataset used, experiments details, the evaluation protocols and the results obtained, in this section.

5.1 Database

The database we would be using is NIH Chest X-ray Dataset [21] of 14 Common Thorax Disease Categories. This dataset comprises 112,120 frontal-view X-ray images of 30,805 unique patients with the text-mined fourteen disease image labels (where each image can have multi-labels), mined from the associated radiological reports using natural language processing. Fourteen common thoracic pathologies include Atelectasis, Consolidation, Infiltration, Pneumothorax, Edema, Emphysema, Fibrosis, Effusion, Pneumonia, Pleural_thickening, Cardiomegaly, Nodule, Mass and Hernia, which is an extension of the 8 common disease patterns listed in their CVPR2017 paper. Note that original radiology reports (associated with these chest x-ray studies) are not meant to be publicly shared for many reasons. The text-mined disease labels are expected to have accuracy >90 %. This dataset is extracted from the clinical PACS database at National Institutes of Health Clinical Center and consists of 60 % of all frontal chest x-rays in the hospital. Therefore, this dataset is significantly more representative to the real patient population distributions and realistic clinical diagnosis challenges, than any previous chest x-ray datasets. Of course, the size of our dataset, in terms of the total numbers of images and thorax disease frequencies, would better facilitate deep neural network training than the previous chest x-ray datasets. The contents of the dataset are:

1. 112,120 frontal-view chest X-ray PNG images in 1024*1024 resolution.

2. Meta data for all images : Image Index, Finding Labels, Follow-up number, Patient ID, Patient Age, Patient Gender, View Position, Original Image Size and Original Image Pixel Spacing.

3. Bounding boxes for 1000 images:Image Index, Finding Label, Bbox[x, y, w, h]

4. Two data split files are provided. Images in the ChestX-ray dataset are divided into these two sets on the patient level. All studies from the same patient will only appear in either training/validation or testing set.

5.2 Experimental setup

The images were resized to 224*224 from the original 1024*1024, and fed into the network for training. Then, the performance was evaluated on the test data given in the test split. The training and validation data was split into 90% training and 10% validation data. Approximately 20% (around 22000) x-ray images are present in the test data for testing the performance on the classification task. But, only 1000 images are present with bounding box labels for checking the performance on the localization task. The parameters for the training process are as follows, the initial learning rate was 0.001, while multiplying it by 0.99 after every epoch. The weight decay rate of 0.0005 was used. We used stochastic batch gradient descent with a batch size of 60. The network was trained for 25 epochs, and the model obtained after the 19th epoch was used for testing (based on validation accuracy on classification task). Inference was performed as described above and in figure 3, and the results obtained according the evaluation protocol used (described below) are reported in the table.

5.3 Evaluation protocol

Different evaluation metrics have been used for the classification and the localization problem.

For the classification problem, we used different thresholds (if score for a classification is above the threshold, it is considered as a positive classification, other a negative classification). We use the thresholds [0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1]. For each threshold value, and for each class, we calculate the true positives (tp), false positives (fp), true negatives (tn) and false negatives (fn). Then, we calculate the true positive rate (tpr) and the false positive rate (fpr) for each class and each threshold value using the equations : $tpr = tp/(tp+fn)$, $fpr = fp/(tn+fp)$ Using the tpr and fpr for each class and different thresholds, the ROC curve is plotted for each class by varying the output threshold probability (probability above which a sample is classified as positive). Then, the area under the ROC curve (AUC) is used as the metric. So, in the results the AUC for each class is calculated separately.

For the localization task, for each input image, a forward pass is done in the network. Since, the labels in the localization task has only one pathology present corresponding to one x-ray, we take that pathology which has

the highest probability according to the final classification score. Then, we take the heatmap corresponding to that pathology and generate the bounding box. For the localization task, to examine the accuracy of computed bounding boxes vs the ground truth bounding boxes, the intersection over union (IoU) is calculated. And then a correct localization is defined if $IOU > T(IoU)$ where $T(Iou) \in 0.1, 0.25, 0.5, 0.75, 0.9$.
For each value of T(IoU), the accuracy and average false positive (AFP) is calculated, and this is done for each class. The results are shown in figure 4.

5.4 Results

The results for the multi-label classification tasks as compared with the baseline (using the VGG model), and the state of the art are given in table 1.

Table 1: Multi-label classification performance (measured by classwise AUC) of our model, baseline given by the original authors (Wang et al) on VGG-net, and the state of the art Chexnet (trained on 121 layers densenet by [18])

Pathologies	Our model	Baseline	Chexnet
Atelectasis	0.6893	0.6281	0.8094
Cardiomegaly	0.7197	0.7084	0.9248
Effusion	0.7646	0.6502	0.8638
Infiltration	0.6563	0.5896	0.7345
Mass	0.6744	0.5103	0.8676
Nodule	0.6459	0.6556	0.7802
Pneumonia	0.5194	0.5100	0.7680
Pneumothorax	0.71974	0.7516	0.8887
Consolidation	0.6150	NA	0.7901
Edema	0.6898	NA	0.8878
Emphysema	0.7128	NA	0.9371
Fibrosis	0.5533	NA	0.8047
Pleural_Thickening	0.6100	NA	0.8062
Hernia	0.5116	NA	0.9164

We see that the performance of our model is better than the baseline given by the original authors (in most of the classes) although the overall idea is pretty much the same, and also the same VGG network is used. This may be attributed because we trained on the new released dataset with 14 classes, whereas the baseline was trained on dataset with 8 classes. The authors have also trained on the dataset with 14 classes, but the results are reported with the resnet network which has much higher complexity than VGG-net and hence the comparison is not very apt. Also there are slight modifications such as not using the whole VGG-net layers (remove the ending 4 conv layers and 1 max-pool layer), so that the size of feature map is not made very small. These slight modifications give a edge in performance. However the performance of the state of the art is much higher than our model, this is because they have used the Densenet network with 121 layers which has a much higher complexity than our network, while also having the disadvantage of high training/test time.

The results for the localization task are given in table 2. We found that the localization results obtained according to the metrics are very nice. This is due to the fact that the number of true negatives are very high, i.e. for each image, the diseases which have been classified as absent and which are actually absent are classified as true negative, they do not actually indicate the success of the localization algorithm, but contribute positively to the accuracy and average false positive. In the paper by Wang et al, they have performance way lower than what we have obtained, and they have not mentioned how they have counted true negative and other measures, so we decided not to include their results because there might be discrepancy in the way of measuring, and would not have been a fair comparison. Qualitative results for the localization task are shown in figure 4.

6 Conclusion and future work

We developed a CNN model for handling the multi-label classification problem. We saw that by using a simple modification of adding a global average pooling layer, we can use the feature maps and weights of the network to generate a likelihood map which identifies those discriminative spatial locations which are used by the CNN to identify the that class. We see that by using this technique, decent performance in weakly supervised object localization can be obtained.
Future work includes experimenting with the loss function (probably with triplet loss to better learn the discriminative features between different pathologies), and using data augmentation to increase the positive examples.

7 References

[1] Himanshu Arora, Nicolas Loeff, David A Forsyth, and Narendra Ahuja. Unsupervised segmentation of objects using efficient learning. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–7. IEEE, 2007.

[2] Loris Bazzani, Alessandra Bergamo, Dragomir Anguelov, and Lorenzo Torresani. Self-taught object localization with deep networks. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–9. IEEE, 2016.

[3] Sima Behpour and Brian D Ziebart. Adversarial methods improve object localization. In *Advances in Neural Information Processing Systems Workshop*, 2016.

[4] Matthew Blaschko, Andrea Vedaldi, and Andrew Zisserman. Simultaneous object detection and ranking with weak supervision. In *Advances in neural information processing systems*, pages 235–243, 2010.

[5] Ondrej Chum and Andrew Zisserman. An exemplar model for learning object classes. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[6] Ramazan Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid. Weakly supervised object localization with multi-fold multiple instance learning. *IEEE transactions on pattern analysis and machine intelligence*, 39(1):189–203, 2017.

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

[8] Thomas Deselaers, Bogdan Alexe, and Vittorio Ferrari. Localizing objects while learning their appearance. In *European conference on computer vision*, pages 452–466. Springer, 2010.

[9] Ali Diba, Vivek Sharma, Rainer Stiefelhagen, and Luc Van Gool. Object discovery by generative adversarial & ranking networks. *arXiv preprint arXiv:1711.08174*, 2017.

[10] Robert Fergus, Pietro Perona, and Andrew Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 2, pages II–II. IEEE, 2003.

[11] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[12] Pulkit Kumar, Monika Grewal, and Muktabh Mayank Srivastava. Boosted cascaded convnets for multilabel classification of thoracic diseases in chest radiographs. *arXiv preprint arXiv:1711.08760*, 2017.

[13] Jianan Li, Xiaodan Liang, Yunchao Wei, Tingfa Xu, Jiashi Feng, and Shuicheng Yan. Perceptual generative adversarial networks for small object detection. In *IEEE CVPR*, 2017.

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

[15] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1717–1724. IEEE, 2014.

[16] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *Proceedings of the IEEE Conference on*

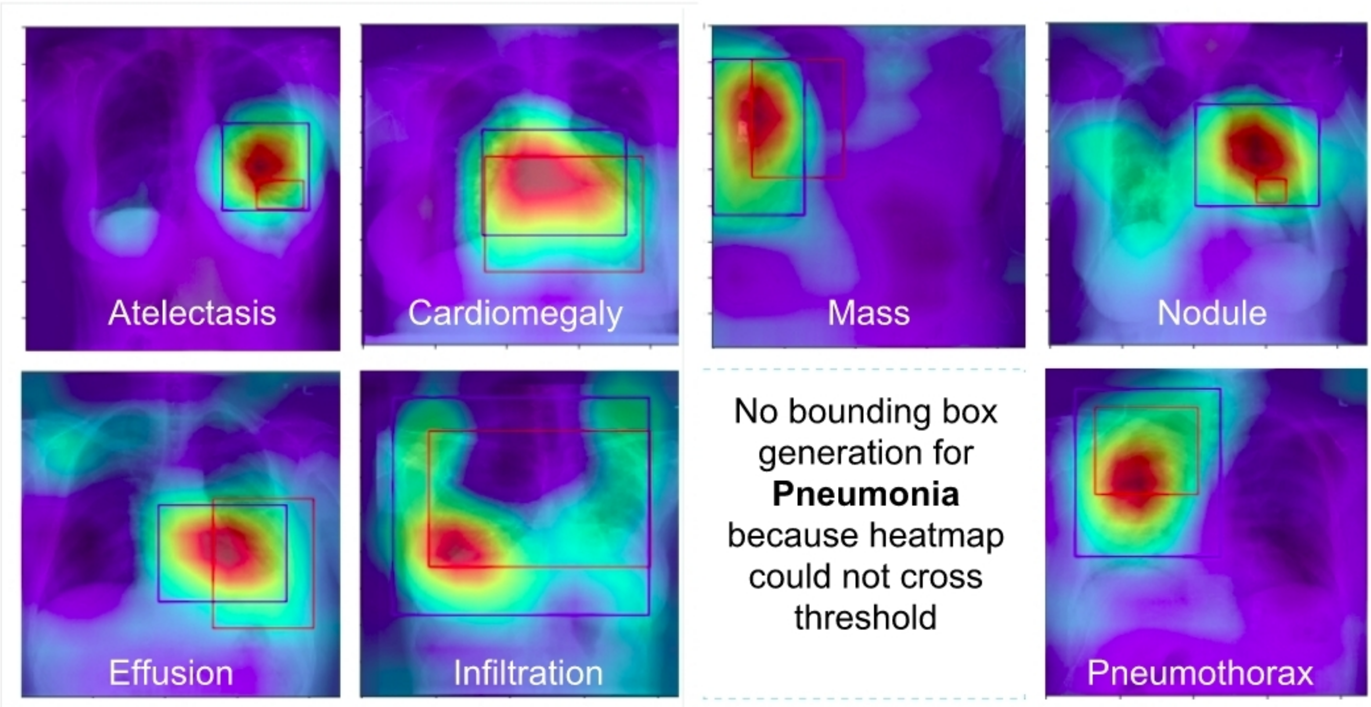
Atelectasis
Cardiomegaly
Mass
Nodule
Effusion
Infiltration
No bounding box generation for **Pneumonia** because heatmap could not cross threshold
Pneumothorax

Figure 4: Qualitative results for the localization task, blue box represent the bounding box generated by us, and the red box represent the ground truth bounding box.

Table 2: Pathology localization accuracy and average false positive (AFP) for each T(IoU) value as mentioned in the evaluation protocol. The diseases are Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, Pneumothorax, Consolidation, Edema, Emphysema, Fibrosis, Pleural_Thickening and Hernia

	Atel.	Card.	Effu.	Infil.	Mass	Nod.	Pneu.	Pneumo.	Conso.	Edema	Emph.	Fibr.	Pleu.	Hernia
T(IoU) = 0.1														
Acc.	0.7488	0.9011	0.7204	0.6670	0.8784	0.8795	0.8659	0.8772	0.9988	0.9954	0.9852	0.9988	0.9931	1.
AFP.	0.1481	0.0228	0.2446	0.2912	0.0591	0.0663	0.	0.0403	0.0011	0.0045	0.0147	0.0011	0.0068	0.
T(IoU) = 0.25														
Acc.	0.7375	0.9	0.7011	0.6534	0.875	0.8795	0.8659	0.8693	0.9988	0.9954	0.9852	0.9988	0.9931	1.
AFP.	0.1592	0.0241	0.2604	0.3017	0.0626	0.0663	0.	0.0486	0.0011	0.0045	0.0147	0.0011	0.0068	0.
T(IoU) = 0.5														
Acc.	0.7318	0.8545	0.6852	0.6477	0.8681	0.8795	0.8659	0.8659	0.9988	0.9954	0.9852	0.9988	0.9931	1.
AFP.	0.1647	0.0737	0.2730	0.3059	0.069	0.0663	0.	0.0522	0.0011	0.0045	0.0147	0.0011	0.0068	0.
T(IoU) = 0.75														
Acc.	0.7318	0.8272	0.6818	0.6454	0.8670	0.8795	0.8659	0.8659	0.9988	0.9954	0.9852	0.9988	0.9931	1.
AFP.	0.1647	0.1012	0.2756	0.3076	0.0706	0.0663	0.	0.0522	0.0011	0.0045	0.0147	0.0011	0.0068	0.
T(IoU) = 0.9														
Acc.	0.7318	0.8272	0.6806	0.6443	0.8670	0.8795	0.8659	0.8659	0.9988	0.9954	0.9852	0.9988	0.9931	1.
AFP.	0.1647	0.1012	0.2765	0.3085	0.0706	0.0663	0.	0.0522	0.0011	0.0045	0.0147	0.0011	0.0068	0.

Computer Vision and Pattern Recognition, pages 685–694, 2015.

[17] Megha Pandey and Svetlana Lazebnik. Scene recognition and weakly supervised object localization with deformable part-based models. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1307–1314. IEEE, 2011.

[18] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

[19] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013.

[20] Hyun Oh Song, Ross Girshick, Stefanie Jegelka, Julien Mairal, Zaid Harchaoui, and Trevor Darrell. On learning to localize objects with minimal supervision. *arXiv preprint arXiv:1403.1024*, 2014.

[21] 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 *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3462–3471. IEEE, 2017.

[22] John Winn and Nebojsa Jojic. Locus: Learning object classes with unsupervised segmentation. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 756–763. IEEE, 2005.

[23] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, pages 2921–2929. IEEE, 2016.