

---

# Weakly Supervised Lesion Localization With Probabilistic-CAM Pooling

---

**Wenwu Ye**

JF Healthcare

wenwu.ye@jfhealthcare.com

**Jin Yao**

JF Healthcare

jin.yao@jfhealthcare.com

**Hui Xue**

JF Healthcare

hui.xue@jfhealthcare.com

**Yi Li**

Greybird Ventures LLC

yil8@uci.edu

## Abstract

Localizing thoracic diseases on chest X-ray plays a critical role in clinical practices such as diagnosis and treatment planning. However, current deep learning based approaches often require strong supervision, e.g. annotated bounding boxes, for training such systems, which is infeasible to harvest in large-scale. We present Probabilistic Class Activation Map (PCAM) pooling, a novel global pooling operation for lesion localization with only image-level supervision. PCAM pooling explicitly leverages the excellent localization ability of CAM [10] during training in a probabilistic fashion. Experiments on the ChestX-ray14 [7] dataset show a ResNet-34 [1] model trained with PCAM pooling outperforms state-of-the-art baselines on both the classification task and the localization task. Visual examination on the probability maps generated by PCAM pooling shows clear and sharp boundaries around lesion regions compared to the localization heatmaps generated by CAM. PCAM pooling is open sourced at <https://github.com/jfhealthcare/Chexpert>.

## 1 Introduction

Computer-aided thoracic disease diagnosis based on chest X-ray has been significantly advanced by deep convolutional neural networks (DCNNs) in recent years [7, 5, 4, 6]. Most of these approaches are formulated as a multi-task binary classification problem, where a CNN is trained to predict the risk probabilities of different thoracic diseases. In clinical practices, visual localization of the lesions on chest X-ray, such as heatmaps or segmentation masks, is also preferred to provide interpretable supports for the classification results. Precise lesion localization often requires training CNNs with strong supervision, such as bounding boxes [7], beyond merely image-level labels. However, accurately annotating lesion locations is difficult, time-consuming, and infeasible to practice in large-scale. For example, one of the largest publicly available chest X-ray datasets ChestX-ray14 [7] contains more than one hundred thousands images with image-level labels, among which only less than one thousand images are further annotated with bounding boxes for benchmarking. Therefore, weakly supervised lesion localization on chest X-ray based on image-level labels remains a challenge but vital problem for computer-aided thoracic disease diagnosis.

The recent work of Class Activation Map (CAM) [10] demonstrates the excellent localization ability of CNNs trained on nature images with only image-level supervision. On chest X-ray images, CAM and its variations have also been used for lesions localization [7, 5, 4, 6]. However, most of these approaches utilize CAM as a post-processing technique to first generate lesion localization heatmaps, then threshold the heatmap scores and generate lesion bounding boxes. We argue that it may be beneficial to leverage CAM even during training given its excellent localization ability.

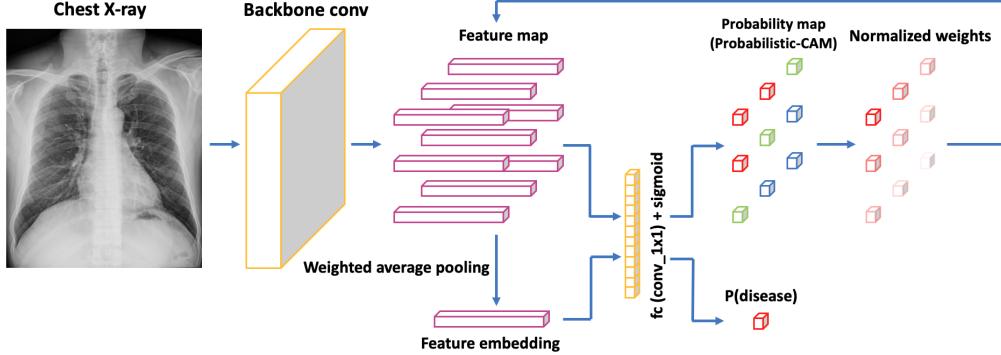


Figure 1: The framework of Probabilistic-CAM (PCAM) pooling.

In this work, we propose a novel and simple extension to the CAM-based framework for lesion localization on chest X-ray with image-level supervision. Specifically, we propose a new global pooling operation that explicitly leverages CAM for localization during training in a probabilistic fashion, namely Probabilistic-CAM (PCAM) pooling. Figure 1 shows the framework of PCAM pooling. A fully convolutional backbone network first processes the input chest X-ray image and generates a feature map. Then, for a particular label of thoracic disease, e.g. ‘‘Pneumonia’’, each feature embedding within the feature map goes through a fully connected (fc) layer implemented as a  $1 \times 1$  convolutional layer and generates the class activation score that monotonically measure the disease likelihood of each embedding. Unlike the standard practice that directly uses the class activation score for localization, we further bound it with the sigmoid function and interpret the output as the disease probability of each embedding. Finally, the output probability map is normalized to attention weights of each embedding, following the multiple-instance learning (MIL) framework [2, 9], which are used to pool the original feature map by weighted average pooling. The pooled embedding goes through the same fc layer introduced above and generates the image-level disease probability for training. During inference time, we directly use the probability map for lesion localization, and apply a simple probability thresholding to obtain disease regions and bounding boxes.

PCAM pooling does not introduce any additional training parameters and is easy to implement. We experiment the efficacy of PCAM pooling for lesion localization with image-level supervision on the ChestX-ray14 [7] dataset. A ResNet-34 [1] model trained with PCAM pooling significantly outperforms the ChestX-ray14 baseline[7] in both the classification task and the localization task. Qualitative visual examination shows the probability maps generated by PCAM pooling tend to have clear and sharp boundaries around lesion regions compared to the typical class activation map.

## 2 Related work

### 2.1 Weakly supervised lesion localization

Various methods have been proposed to localize lesions on chest X-ray through image-level supervision [7, 5, 4, 9, 6]. Wang et al. [7] introduce the ChestX-ray14 dataset, together with a baseline for evaluating weakly supervised lesion localization. Instead of the typical global average pooling, Wang et al. use the Log-Sum-Exp (LSE) pooling [3] to encourage model training focusing more on the discriminative regions of the chest X-ray. Sedai et al. [4] utilize the intermediate feature maps from CNN layers at different scales together with learned layer reference weights to improve the localization performance of small lesions. Tang et al. [5] combine CNN with attention-guided curriculum learning to gradually learn distinctive convolutional features for localization. Most of these approaches utilize the standard CAM technique to localize lesions, where our proposed PCAM pooling serves as an extension to the standard CAM to improve lesion localization with image-level supervision.

Table 1: Different types of global pooling.

Pooling type	Formulation
Average	$x_c = \sum_{i,j}^{H,W} w_{c,i,j} X_{c,i,j}, w_{c,i,j} = \frac{1}{H*W}$
Linear [8]	$x_c = \sum_{i,j}^{H,W} w_{c,i,j} X_{c,i,j}, w_{c,i,j} = \frac{X_{c,i,j}}{\sum_{i,j}^{H,W} X_{c,i,j}}$
Exponential [8]	$x_c = \sum_{i,j}^{H,W} w_{c,i,j} X_{c,i,j}, w_{c,i,j} = \frac{\exp(X_{c,i,j})}{\sum_{i,j}^{H,W} \exp(X_{c,i,j})}$
LSE [3]	$x_c = \frac{1}{\gamma} \log \left[ \frac{1}{H*W} \sum_{i,j}^{H,W} \exp(\gamma X_{c,i,j}) \right]$
LSE-LBA [9]	$s = \frac{1}{\gamma_0 + \exp(\beta)} \log \left[ \frac{1}{H*W} \sum_{i,j}^{H,W} \exp [(\gamma_0 + \exp(\beta)) S_{i,j}] \right]$
Attention [2]	$x = \sum_{i,j}^{H,W} w_{i,j} X_{i,j}, w_{i,j} = \frac{\exp[\mathbf{w}^\top \tanh(\mathbf{V} X_{i,j})]}{\sum_{i,j}^{H,W} \exp[\mathbf{w}^\top \tanh(\mathbf{V} X_{i,j})]}$

## 2.2 Global pooling

Global average pooling is arguably the most widely used global pooling operation for image classification. While it is less prone to overfitting as the network tries to identify all discriminative regions of an object in natural images [10], it may also fail to highlight the most discriminative regions within a chest X-ray, given most chest X-ray images share the same anatomic structures and may only differ in fine-grained lesion regions. Therefore, different global pooling operations have also been used to analyze chest X-rays. For example, Wang et al. [7] uses LSE pooling, which can be viewed as an adjustable operation between max pooling and average pooling by controlling the hyper-parameter  $\gamma$ .

We summarize the mathematical differences and correlations between different global pooling operations in Table 1. Particularly,  $X$  with shape  $(C, H, W)$  denotes the feature map from the last convolutional layer of a CNN.  $x$  denotes the feature embedding of length  $C$  after global pooling.  $C$  denotes the channel dimension,  $H$  and  $W$  denote the height and width of the feature map.

We can see that, global pooling operations differ mainly in the ways to compute the weights for feature map averaging. Note that, most of the pooling is performed for each channel  $X_{c,i,j}$  independently, except for Attention pooling [2] that computes the attention weight at embedding level  $X_{i,j}$  with extra trainable parameters, i.e.  $\mathbf{V}, \mathbf{w}$  in Table 1. All the channels within the same embedding share the same weight. The basic assumption of Attention pooling follows the multiple-instance learning (MIL) framework, that treats each embedding as an instance, and the chest X-ray as a bag of instances is positive, e.g. certain thoracic disease, as long as one of the instance is positive. PCAM pooling follows the same MIL framework, but uses a different method to compute the attention weight for each embedding based on the localization ability of CAM without introducing extra trainable parameters. LSE-LBA [9] also falls within the MIL framework, but it performs the pooling operation on the saliency map  $S$  of shape  $(H, W)$  instead of the feature map  $X$  to obtain a saliency score  $s$  for training.

## 3 Probabilistic-CAM (PCAM) Pooling

The main idea of PCAM pooling is to explicitly leverage the localization ability of CAM [10] through the global pooling operation during training. Using the same notation from Section 2.2, given a fully convolutional network trained for multi-task binary classification, the class activation map of a particular thoracic disease is given by  $\{s_{i,j} = \mathbf{w}^\top X_{i,j} + b | i, j \in H, W\}$ .  $X_{i,j}$  is the feature embedding of length  $C$  at the position  $(i, j)$  of a feature map  $X$  with shape  $(C, H, W)$  from the last convolutional layer.  $\mathbf{w}, b$  are the weights and bias of the last fc layer for binary classification. In other words,  $s_{i,j}$  is the logit before sigmoid function under the binary classification setting.  $s_{i,j}$  monotonically measures the disease likelihood of  $X_{i,j}$ , and is used to generate the localization heatmap after the model is trained in the standard CAM framework.

PCAM pooling utilizes  $s_{i,j}$  to guide lesion localization during training through the global pooling operation under the MIL framework [2]. The MIL framework assumes the chest X-ray as a bag of embeddings is positive, as long as one of the embedding is positive. To measure each embedding's contribution to the whole bag, the MIL framework assigns normalized attention weights to each embedding for weighted global average pooling [2]. Because the numerical range of  $s_{i,j}$  is unbounded, i.e.  $s_{i,j} \in (-\infty, +\infty)$  in theory, it's neither interpretable nor directly applicable to compute the

attention weights. Therefore, we further bound  $s_{i,j}$  with the sigmoid function,  $p_{i,j} = \text{sigmoid}(s_{i,j})$ , and normalize  $p_{i,j}$  as the attention weights. In summary, PCAM pooling can be formulated as

$$x = \sum_{i,j}^{H,W} w_{i,j} X_{i,j}, \quad w_{i,j} = \frac{\text{sigmoid}(\mathbf{w}^\top X_{i,j} + b)}{\sum_{i,j}^{H,W} \text{sigmoid}(\mathbf{w}^\top X_{i,j} + b)} \quad (1)$$

where  $w_{i,j}$  is the attention weight for  $X_{i,j}$  and  $x$  is the pooled feature embedding which goes through the same fc layer for final image level classification.

During the inference time, we interpret the sigmoid-bounded  $p_{i,j}$  as the probability of embedding  $X_{i,j}$  being positive, thus named Probabilistic-CAM, and we use the probability map  $\{p_{i,j}|i, j \in H, W\}$  as the localization heatmap. We use a simple probability thresholding on the probability map to obtain regions of interest. In comparison, because  $s_{i,j}$  is unbounded with different numerical ranges in different chest X-ray images, it is usually normalized into  $[0, 255]$  within each image and thresholded with some ad-hoc ranges, e.g.  $[60, 180]$  in [7], to generate regions of interest. We show in Section 4.3 section that PCAM pooling generates localization heatmaps with better visual quality around lesion boundary regions compared to the standard CAM.

## 4 Experiments

### 4.1 Dataset and experiments setup

We evaluate lesion localization with image-level supervision on the ChestX-ray14 [7] dataset, which contains 112,120 frontal-view chest X-ray images with 14 thoracic disease labels. 8 out of the 14 diseases are further annotated with 984 bounding boxes. We randomly split the official train\_valid set into 75% for training and 25% for validation. On the official test set, we evaluate the classification task on the 14 diseases and the localization task on the 8 diseases that have bounding boxes. Note that the 984 bounding boxes are not used for training.

We use ResNet-34 [1] as the backbone network and process the input images on the original  $1024 \times 1024$  scale following [7]. The network is trained with a batch size of 36 and a learning rate of  $1e^{-4}$  for 10 epoches. We balance the binary cross entropy loss of positive and negative samples within each batch following [7]. For the localization task, we first apply a probability of 0.9 to threshold the probability maps from PCAM pooling, then generate bounding boxes that cover the isolated regions in the binary masks. To compare the visual quality of localization heatmaps from PCAM pooling with previous methods, we also train a ResNet-34 with LSE pooling following [7]. We normalize the class activation maps from LSE pooling into  $[0, 255]$  for each image individually, and then apply a threshold of 180 following [7]. Note that the performances of LSE pooling reported in Table 2 and Table 3 are from Wang et al. [7].

### 4.2 Classification task

Table 2 shows the Area Under the receiver operating characteristic Curves (AUCs) of the classification task on the 14 thoracic diseases from the ChestX-ray14 official test set. PCAM pooling outperforms most of the other state-of-the-art methods including the baseline reported in ChestX-ray14 [7]. We suspect explicitly utilizing CAM for localization during training may also benefit the classification task. Note the results from AGCL [5] are obtained by training with additional severity-level information.

### 4.3 Localization task

Table 3 shows the localization accuracies and the average false positives of the localization task on the 8 thoracic diseases that have bounding boxes. We use Intersection over the predicted B-Box area ratio (IoBB) to measure the overlap between predicted bounding boxes and ground truth bounding boxes annotated by radiologists following [7]. A correct localization is defined as at least one predicted bounding box is overlapped with the ground truth bounding box with  $\text{IoBB} > 0.5$  [7]. PCAM pooling outperforms the baseline localization accuracy [7] by a significant margin on all of the diseases, demonstrating its efficacy in weakly supervised lesion localization.

Figure 2 shows a few selected examples of the probability maps generated by PCAM pooling and the class activation maps generated by LSE pooling together with the predicted bounding boxes.

Table 2: Classification AUCs of the 14 diseases on the ChestX-ray14 test set.

Method	LSE [7]	LSE-LBA [9]	AGCL [5]	ChestNet [6]	PCAM pooling
Atelectasis	0.700	0.733	0.756	0.743	<b>0.772</b>
Cardiomegaly	0.810	0.856	<b>0.887</b>	0.875	0.864
Effusion	0.759	0.806	0.819	0.811	<b>0.825</b>
Infiltration	0.661	0.673	0.689	0.677	<b>0.694</b>
Mass	0.693	0.777	<b>0.814</b>	0.783	0.813
Nodule	0.669	0.718	0.755	0.698	<b>0.783</b>
Pneumonia	0.658	0.684	<b>0.729</b>	0.696	0.721
Pneumothorax	0.799	0.805	0.850	0.810	<b>0.868</b>
Consolidation	0.703	0.711	0.728	0.726	<b>0.732</b>
Edema	0.805	0.806	<b>0.848</b>	0.833	0.833
Emphysema	0.833	0.842	0.908	0.822	<b>0.931</b>
Fibrosis	0.786	0.743	0.818	0.804	<b>0.819</b>
Pleural Thickening	0.684	0.724	0.765	0.751	<b>0.788</b>
Hernia	0.872	0.775	0.875	<b>0.900</b>	0.784

Table 3: Localization accuracies and average false positives of the 8 diseases that have bounding boxes on the ChestX-ray14 test set.

Method	AT	CM	EF	IF	MS	ND	PMN	PMT
Localization accuracy, IoBB > 0.5								
LSE-baseline [7]	0.2833	0.8767	0.3333	0.4227	0.1411	0.0126	0.3833	0.1836
PCAM pooling	<b>0.3500</b>	<b>0.9657</b>	<b>0.5359</b>	<b>0.7642</b>	<b>0.4118</b>	<b>0.0759</b>	<b>0.7667</b>	<b>0.1939</b>
Average false positive								
LSE-baseline [7]	1.020	0.563	<b>0.927</b>	<b>0.659</b>	<b>0.694</b>	<b>0.619</b>	<b>1.013</b>	<b>0.529</b>
PCAM pooling	<b>0.867</b>	<b>0.021</b>	1.137	1.805	1.000	1.228	1.200	1.684

AT: Atelectasis, CM: Cardiomegaly, EF: Effusion, IF: Infiltration, MS: Mass, ND: Nodule, PMN: Pneumonia, PMT: Pneumothorax.

Compared to the class activation maps, the probability maps are visually more clear with sharp boundaries around lesion regions. We attribute the improved visual quality to the probabilistic interpretation of the sigmoid-bounded class activation map and explicitly using it for training with global pooling.

We notice the probability maps generated by PCAM pooling tend to enlarge regions of interest in general than class activation maps from LSE pooling, especially when the ground truth regions are small, such as “Nodule” in Figure 2. This may explain the fact that PCAM pooling has relatively larger average false positives than CAM with LSE pooling.

## 5 Conclusion

In this work, we present Probabilistic-CAM (PCAM) pooling, a new global pooling operation to explicitly leverage the localization ability of CAM for training. PCAM pooling is easy to implement and does not introduce any additional training parameters. Experiments of weakly supervised lesion localization on the ChestX-ray14 [7] dataset demonstrate its efficacy in improving both the classification task and the localization task compared to several state-of-the-art baselines. Visual inspection of the probability maps generated by PCAM pooling shows clear and sharp boundaries around lesion regions compared to the standard class activation maps.

Currently, PCAM pooling tends to generate localization maps that enlarge regions of interest, which may increase false positives especially for small lesions. We are working on reducing this effect as our future direction.

## References

- [1] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

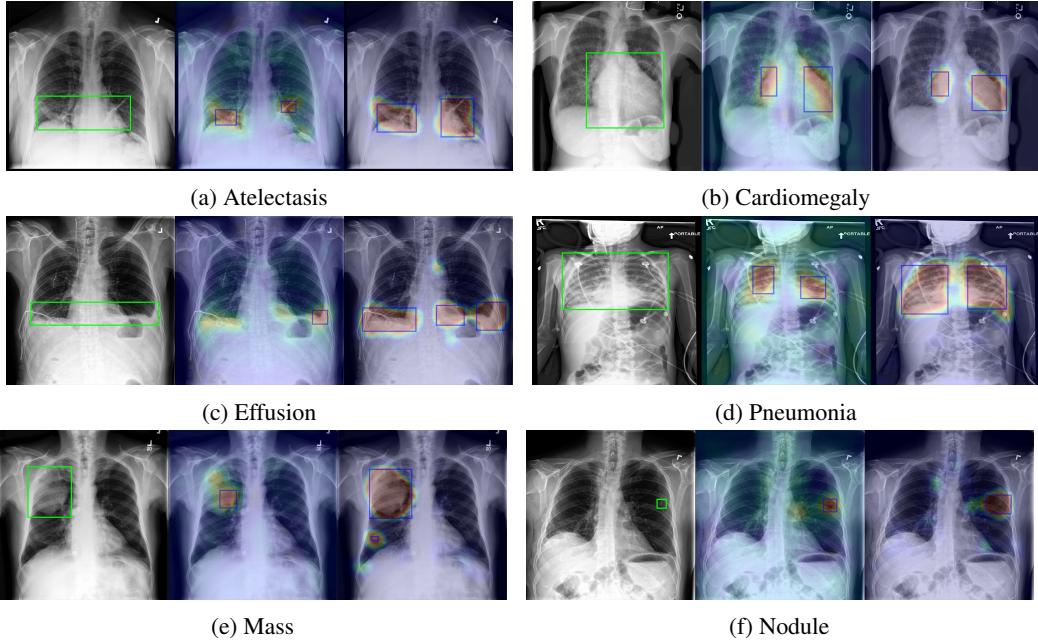


Figure 2: Selected samples of localization heatmaps and their bounding boxes generated by LSE pooling and PCAM pooling on the test set of ChestX-ray14 [7]. In each subfigure, the left panel is the original chest X-ray with the ground truth bounding boxes (green), the middle panel is the class activation map and predicted bounding boxes (blue) by LSE pooling, the right panel is the probability map and predicted bounding boxes (blue) by PCAM pooling.

- [2] M. Ilse, J. M. Tomczak, and M. Welling. Attention-based deep multiple instance learning. *arXiv preprint arXiv:1802.04712*, 2018.
- [3] P. O. Pinheiro and R. Collobert. From image-level to pixel-level labeling with convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1713–1721, 2015.
- [4] S. Sedai, D. Mahapatra, Z. Ge, R. Chakravorty, and R. Garnavi. Deep multiscale convolutional feature learning for weakly supervised localization of chest pathologies in x-ray images. In *International Workshop on Machine Learning in Medical Imaging*, pages 267–275. Springer, 2018.
- [5] Y. Tang, X. Wang, A. P. Harrison, L. Lu, J. Xiao, and R. M. Summers. Attention-guided curriculum learning for weakly supervised classification and localization of thoracic diseases on chest radiographs. In *International Workshop on Machine Learning in Medical Imaging*, pages 249–258. Springer, 2018.
- [6] H. Wang and Y. Xia. Chestnet: A deep neural network for classification of thoracic diseases on chest radiography. *arXiv preprint arXiv:1807.03058*, 2018.
- [7] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.
- [8] Y. Wang, J. Li, and F. Metze. A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 31–35. IEEE, 2019.
- [9] L. Yao, J. Prosky, E. Poblenz, B. Covington, and K. Lyman. Weakly supervised medical diagnosis and localization from multiple resolutions. *arXiv preprint arXiv:1803.07703*, 2018.
- [10] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.