Thoracic Disease Identification and Localization using Distance Learning and Region Verification

Cheng Zhang[1]
zhang.7804@osu.edu

Francine Chen[2]
francine@acm.org

Yan-Ying Chen[2]
yanying@gmail.com

[1] The Ohio State University
Columbus, OH, USA

[2] FX Palo Alto Laboratory
Palo Alto, CA, USA

Abstract

The identification and localization of diseases in medical images using deep learning models have recently attracted significant interest. Existing methods only consider training the networks with each image independently and most leverage an activation map for disease localization. In this paper, we propose an alternative approach that learns discriminative features among triplets of images and cyclically trains on region features to verify whether attentive regions contain information indicative of a disease. Concretely, we adapt a distance learning framework for multi-label disease classification to differentiate subtle disease features. Additionally, we feed back the features of the predicted class-specific regions to a separate classifier during training to better verify the localized diseases. Our model can achieve state-of-the-art classification performance on the challenging ChestX-ray14 dataset, and our ablation studies indicate that both distance learning and region verification contribute to overall classification performance. Moreover, the distance learning and region verification modules can capture essential information for better localization than baseline models without these modules.

1 Introduction

Radiography has been widely adopted for detecting a number of thoracic diseases. However, detecting diseases in X-ray images requires the expert knowledge of radiologists, who are overburdened and often must quickly review each image. Further, the location of an identified disease is generally not annotated and so may be unclear to another doctor reviewing the X-ray. Several datasets have been released for disease classification from chest X-ray images, including [3, 12, 14, 29], and the use of deep learning models in combination with the datasets has resulted in much progress [4, 5, 16, 23, 29, 30]. Nevertheless, the identification and localization of thoracic diseases are still challenging due to subtle inter-disease differences and large intra-disease variations across different subjects and regions.

The only chest X-ray dataset with disease bounding boxes, ChestX-ray14 [29], has boxes only for the test partition. Under this constraint, it is difficult to apply supervised learning
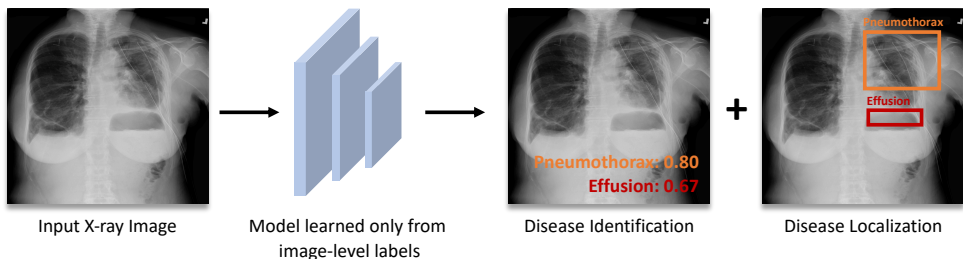
Figure 1: **Overview of thoracic disease identification and localization** trained on chest X-rays and image-level disease labels. Given a test image, the model predicts how likely the diseases occur in the image and localizes their corresponding bounding boxes. Multiple diseases could co-exist in an image.

to disease localization, and two popular weakly supervised approaches have been proposed for disease identification and localization: CAM-based (class activation map) [34] and MIL-based (multi-instance learning). Both embed an X-ray image using a pre-trained image embedding network such as ResNet [8] or DenseNet [11]. Under CAM-based, the computed features are used to train a multi-label classifier on disease labels; and localization is performed based on the class activations in the embedding network [22, 29]. In MIL-based, a grid is formed from the embedded features and each element of the grid is classified as to which, if any, diseases occur, indicating which diseases occur in the image; and localization is performed by combining the grid classifications [16, 17].

In this paper, we focus on multi-label, weakly supervised thoracic disease identification (we will use the terms *identification* and *classification* interchangeably) and localization in chest X-rays using a weakly supervised learning approach (see Figure 1) . Different from existing methods that learn a neural network model on each image independently, we leverage distance learning [1, 10, 13, 26, 32, 33] to learn sufficient feature representations to tackle multi-label disease classification. In particular, we exploit triplets of images as the inputs to drive the similarity metric to be small for the pairs of images with similar diseases, and large for the pairs with different ones. Furthermore, we propose a region verification module that feeds back the class-specific attentive regions to verify those regions important to disease classification. The size of the regions vary, in contrast to the full images used with previous CAM-based approaches and the smaller contexts from the grids used in the MIL-based approaches. Finally, with the well-learned disease features, our model can be used for disease localization by incorporating weakly-supervised object detection methods such as CAM.

Evaluation on the ChestX-ray14 dataset shows that our model achieves state-of-the-art classification performance and consistently outperforms CAM and MIL baselines for multiple diseases. Ablation studies indicate that both distance learning and region verification help classification performance, with the strongest contribution from region verification. Moreover, our experiments also indicate that both modules improve disease localization over baseline models without these modules, strengthening the support to the improvement on classification since the diseases are better localized by verified regional features. In conclusion, our contributions include: (1) design an end-to-end framework for training multi-label disease classification and weakly supervised localization simultaneously; (2) extend distance learning to multi-disease medical images; (3) propose region verification to align disease classification of a whole image and the local context surrounding the disease.

2 Related Work

Disease identification. Recently, significant progress on X-ray classification and detection has been made using deep neural networks. Bar *et al*. [2] presented an early examination of the strength of deep learning approaches for pathology detection in chest radiographs. A number of works developed thoracic disease classification models on the more comprehensive ChestX-ray14 dataset [4, 7, 16, 17, 22, 23, 29]. Wang *et al*. [29] and Rajpurkar *et al*. [22, 23] explored the use of deep learning models for disease classification. Li *et al*. [16] proposed to unify the training of image-level and box-level labels in one framework with a customized MIL loss where disease classification is performed on a grid over the image. Several works applied an attention mechanism to chest X-ray analysis [4, 7, 21]. For example, Guan *et al*. [7] designed an attention guided two-branch network for thorax disease classification, which helps amplify the *high activation* regions. Cai *et al*. [4] presented an attention mining strategy to improve the model's sensitivity or saliency to disease patterns. Liu *et al*. [17] proposed an alternative method for computing attention based on the difference between the input image and an image without any disease (a "negative image"), but did not employ the pair in distance learning. Different from the above methods that only consider single/pair-level image information and class-agnostic attention features, we take advantages of triplets of images with *distance learning* for *multi-label* disease classification and cyclically train [20] the networks with *class-specific* attentive region features.

Weakly-supervised disease localization. Fully supervised detection methods have achieved great success in identifying objects when trained on a vast number of bounding box annotations [6, 18, 24, 25]. However, such annotations are very expensive to create manually for medical images, requiring busy radiologists to do the annotation. Therefore, weakly supervised approaches to object localization using disease labels for an image are commonly used. One popular approach when classification is performed over the full image is to estimate disease locations based on class activation maps (CAM) [34] or gradient-weighted class activation maps (Grad-CAM) [27]. Zhou *et al*. [34] utilized a global average pooling layer for neural networks to generate CAM that are used to localize objects. The use of CAM-based methods was employed by [4, 7, 9, 16, 22, 23, 29]. Often, the thresholded CAM activations are referred to as "attention" or "attentive regions". In another approach employing MIL as used by Li *et al*. [16] and Liu *et al*. [17], the cells of the image grid for which diseases are predicted are combined to predict disease locations. Most prior works leverage attentive activation maps to localize disease regions while neglecting the domain gap between classification and localization results. We propose region verification which feeds back CAM-based region features to a local classifier to verify the disease localized by CAM. Our region verification module is class-specific and could be also incorporated with other weakly-supervised object localization methods.

Distance learning. Distance (metric) learning (DML) generally works with two types of data: *pair-wise* with must-link and cannot-link constraints, and *triplet* constraints that contains a similar pair and a dissimilar pair. The triplet loss with semi-hard mining was introduced in [26] to compute image embeddings for identifying faces and extended to improve performance or computation [1, 10]. While earlier distance learning tasks focused on re-identification [19, 28, 33] of one class of object per image, we investigate how distance learning can be extended to multi-label disease images. Instead of directly using the disease features extracted by the pre-trained deep model, we present a first attempt to employ triplet learning for a multi-label disease classification task.
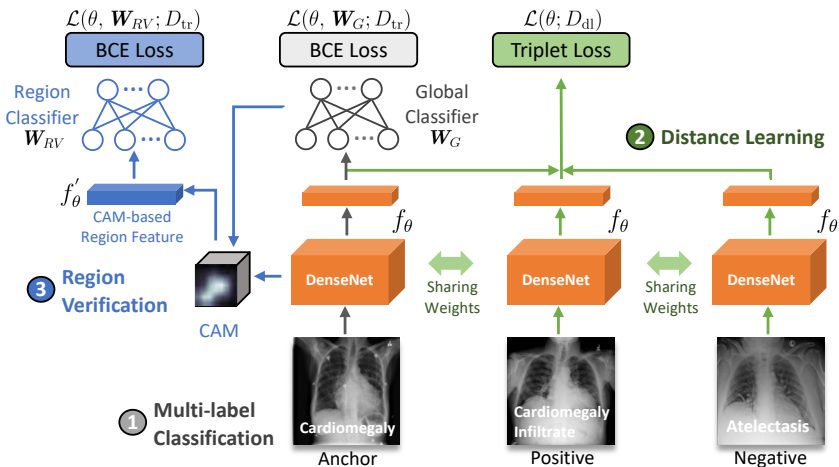
Figure 2: **Framework of disease identification and localization.** Our model takes triplets of images as the inputs, *i.e.*, anchor, positive, and negative for distance learning, and feeds back the CAM-based attentive feature to the region classifier for disease verification. The pipeline is trained end-to-end.

3 Approach

An overview of our proposed training framework is shown in Figure 2. Our model takes triplets of images as the input and consists of three parts: (1) multi-label classification of the anchor image with a conventionally trained binary cross entropy (BCE) loss (2) distance learning with triplet loss and multi-label hard example mining and (3) region verification of attentive local areas with BCE loss. We jointly train the three loss functions end-to-end. We next describe each component.

3.1 Multi-label classification

A disease identification model takes an input X-ray image $\boldsymbol{x}$ and predicts a confidence score for each disease c by

$$p(y_c|\boldsymbol{x}) = \sigma(\boldsymbol{w}_c^\top f_{\boldsymbol{\theta}}(\boldsymbol{x})), \tag{1}$$

where $f_{\boldsymbol{\theta}}(\cdot)$ is an embedding network parameterized by $\boldsymbol{\theta}$, and $\boldsymbol{w}_c$ is the linear classifier of disease type c. Here we formulate disease identification as a multi-label classification problem. Given a training set $D_{\text{tr}} = \{\boldsymbol{x}_n, \boldsymbol{y}_n\}_{n=1}^N$, in which $\boldsymbol{x}_n$ is the input image to be classified and $\boldsymbol{y}_n \in \Delta^C$ is a vector on the C-dimensional simplex. C is the size of the disease types. We train the embedding network using empirical risk minimization (ERM) with BCE loss:

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{W}_G; D_{\text{tr}}) = -\sum_n \sum_c \boldsymbol{y}_{n,c} \log p(y_{n,c}|\boldsymbol{x}_n) + (1 - \boldsymbol{y}_{n,c}) \log(1 - p(y_{n,c}|\boldsymbol{x}_n)), \tag{2}$$

where the global branch classifier $\boldsymbol{W}_G = \{\boldsymbol{w}_c\}_{c=1}^C$ and stochastic gradient descent (SGD) optimization is applied with uniformly sampled instances from D_{tr}.

3.2   Multi-label distance learning with triplet loss

We adapt a multi-label classification system to use a triplet loss to encourage images with at least one common disease to be close together in feature space, and images without any common disease to be far apart in feature space. A triplet loss minimizes the distance between the image to be classified, or *anchor*, and another image of the same type. We examined two cases as *positive* examples: (1) only images with an identical set of labels and (2) images from case (1) and also images with partial label matches. *Negative* examples were defined as images for which the intersection of disease labels with the anchor is null. Since small batches may not contain exact matches for some label combinations, we also proposed to pre-compute similarity on a randomly selected subset of exact match, partial match, or no match images, rather than by batch during training, leading to more efficient training.

Specifically, we consider a distance (triplet) learning constraint set D_{dl} given as follows:

$$D_{dl} = \{(\boldsymbol{x}_i, \boldsymbol{x}_i^+, \boldsymbol{x}_i^-) | (\boldsymbol{x}_i, \boldsymbol{x}_i^+) \in \mathcal{P}; (\boldsymbol{x}_i, \boldsymbol{x}_i^-) \in \mathcal{N}\}, \tag{3}$$

where $i = 1, 2, \ldots, M$ (M is the cardinality of the entire triplet set). $\mathcal{P}$ contains positive pairs and $\mathcal{N}$ includes negative pairs. We denote the similarity of two samples as ℓ_2 distance between the feature embeddings and define the hinge loss for one triplet given the anchor $\boldsymbol{x}_i$:

$$l(\boldsymbol{x}_i, \boldsymbol{x}_i^+, \boldsymbol{x}_i^-) = [\|f_{\boldsymbol{\theta}}(\boldsymbol{x}_i), f_{\boldsymbol{\theta}}(\boldsymbol{x}_i^+)\| - \|f_{\boldsymbol{\theta}}(\boldsymbol{x}_i), f_{\boldsymbol{\theta}}(\boldsymbol{x}_i^-)\| + m]_+, \tag{4}$$

where m is a margin (0.5 in the experiments) that is enforced between positive and negative pairs. Therefore the triplet loss is minimized over all possible triplets in the distance learning set D_{dl}, which can be computed as:

$$\mathcal{L}(\boldsymbol{\theta}; D_{dl}) = \sum_{(\boldsymbol{x}_i, \boldsymbol{x}_i^+, \boldsymbol{x}_i^-) \in D_{dl}} l(\boldsymbol{x}_i, \boldsymbol{x}_i^+, \boldsymbol{x}_i^-). \tag{5}$$

Hard example mining. The distance learning constraint set D_{dl} is constructed by hard example mining, *i.e.*, selection of semi-hard training examples relative to the anchor image, to make training a model more effective [26]. To identify harder examples for training, we sort the set of positive images and the set of negative images for an anchor image by perceptual similarity to the anchor using a perceptual hash [17, 31]. To reduce computation prior to sorting for identifying hard examples for an anchor, a randomly selected set of 1000 negative examples, and a randomly selected set of up to 500 positive examples were identified for each anchor. To give preference to exact matches when partial matches are used, we set 25% of the positive examples to be partial matches, while the remainder are exact matches. The examples for each anchor are randomly selected with a bias towards easier examples initially and the hardest examples after 10 epochs.

3.3   Region verification

Intuitively, if the disease location is predicted incorrectly, the classification will likely be incorrect. To align disease localization and classification, we propose to feed the attentive region features into another region classifier for cyclical training, namely *region verification* (RV). In the existing single-class CAM approach, weighted activations for a single class are used for localization. In contrast, in our multi-label setting, there can be multiple classes. To handle this, we compute class-dependent activations by projecting back the weights of the

output layer separately for each class. We then merge the attentive regions of all identified classes (labeled for training, predicted for testing) in each image and zero out the activations of non-attentive regions to use as features in region verification. This also contrasts with related work that uses class-independent, high activation regions.

As shown in Figure 2, we use CAM [34] to identify the region, where the activations at the last layer of DenseNet are extracted with the weights of global image classifier:

$$M_c(a,b) = \sum_k w_c^k \cdot g_k(a,b), \tag{6}$$

where w_c^k is the scalar weight of disease class c for feature k. We use $g_k(a,b)$ to denote a value of 2-dimensional spatial point (a,b) with feature k in map g_k from the final convolutional layer of DenseNet-121, where the sizes of k and g_k are 1024 and $1024 \times 7 \times 7$, respectively.

To extract attentive region features, our class-specific activation map considers all feature maps associated to only "positive" diseases (labels) in the anchor image:

$$M(a,b) = \sum_c \mathbb{1}[y_c = 1] \cdot M_c(a,b). \tag{7}$$

Next, we normalize the activation map to [0, 1] over the difference between the maximum and minumum value over all features. The locations of feature activations that are greater than a fixed threshold (we empirically set to 0.8 in the experiments) are identified. Then a rectangular bounding box around the thresholded activations is computed and features outside the bounding box are zeroed. The resulting features are fed into the region classifier, which is a separate classifier randomly initialized. Finally, we optimize the region classifier and the shared embedding network by using a binary cross entropy loss as follows:

$$\mathcal{L}(\boldsymbol{\theta}, W_{RV}; D_{tr}) = -\sum_n \sum_c y_{n,c} \log \sigma(v_c^\top f'_{\boldsymbol{\theta}}(x)) + (1 - y_{n,c}) \log(1 - \sigma(v_c^\top f'_{\boldsymbol{\theta}}(x))), \tag{8}$$

where $W_{RV} = \{v_c\}_{c=1}^C$ is the region verification classifier and $f'_{\boldsymbol{\theta}}$ is the distilled CAM-based attentive region feature.

3.4   Joint learning and inference

We jointly optimize the parameters of the model. The gradients are computed from three losses. The main loss $\mathcal{L}(\boldsymbol{\theta}, W_G; D_{tr})$ refers to binary cross entropy loss associated with the whole image prediction from Equation 2. We backpropagate this loss to optimize the parameters of the feature embedding network and global classifier W_G. The region verification loss $\mathcal{L}(\boldsymbol{\theta}, W_{RV}; D_{tr})$ is another binary cross entropy loss for disease class predictions from attentive region features from Equation 8. We use this loss to optimize the shared feature embedding parameters $\boldsymbol{\theta}$ and region verification classifier W_{RV}. By doing so, we further improve the model's ability to capture intra-image subtle disease features. Based on Equation 5, $\mathcal{L}(\boldsymbol{\theta}; D_{dl})$ is the triplet loss during learning of inter-image relationships from distance learning set D_{dl}. We obtain the final loss $\mathcal{L}_{total}$ by adding the three losses together:

$$\mathcal{L}_{total}(\boldsymbol{\theta}, W_G, W_{RV}; D_{tr}) = \mathcal{L}(\boldsymbol{\theta}, W_G; D_{tr}) + \mathcal{L}(\boldsymbol{\theta}, W_{RV}; D_{tr}) + \mathcal{L}(\boldsymbol{\theta}; D_{dl}) \tag{9}$$

Disease identification. We average the decision values by considering both global and region classifiers to compute the confidence score for each disease:

$$p_{total}(y_c|x) = \sigma(w_c^\top f_{\boldsymbol{\theta}}(x) + v_c^\top f'_{\boldsymbol{\theta}}(x)). \tag{10}$$

Disease localization. We average the weights of the global and region classifiers to compute the activation map for one disease c:

$$M_c(a,b) = \sum_k \frac{1}{2}(w_c^k + v_c^k) \cdot g_k(a,b). \tag{11}$$

We normalize the activation map $M_c(a,b)$ and apply a selected threshold to generate the predicted bounding box. Details and qualitative results can be found in the experiments.

4 Experiments

4.1 Setup

Dataset. We evaluate our approach on the NIH ChestX-ray14 dataset [29], which consists of 112,120 frontal-view X-ray images of 30,805 patients with 14 disease labels and each image can have multi-labels, and we follow its official train/test split. In a subset of the test set used for evaluation of the localization task, 880 images are labeled with 984 bounding boxes for 8 diseases by board-certified radiologists. We reserved 10% of the images from the training set as validation based on patient ID, thus insuring that images of an individual patient are present in only one of the train, validation or test sets. Consequently, we split data into 77,821/8,703/25,596 images for training/validation/testing.

Evaluation metrics. We follow [16, 17, 29] to evaluate our approach. For disease classification, we use AUC scores (the area under the ROC curve) to measure the performance. For disease localization, we evaluate the detected regions against the ground truth (GT) bounding boxes, using accuracy for a given intersection over union ratio (IoU). The localization results are only calculated for those 8 disease types with GT provided. A correct localization is defined when IoU > T, where T is the threshold.

Implementation details. We resize the original 3-channel X-rays from 1024×1024 to 224×224 for faster processing. We apply a 5 degree random rotation and horizontal flipping during training for data augmentation. The DenseNet-121 [11] model, pretrained with ImageNet, is used as the backbone because of its better performance on disease classification as shown in [22]. For all models, we train for at most 30 epochs using SGD with Adam [15]. The initial learning rate is 10^{-3}, which is divided by 10 after E epochs. We tune E on the validation set and choose the best model via validation performance. For disease localization, since no box annotations are available for validation, we apply the same threshold on CAM for each disease type using 10-fold cross-validation of the test data to predict bounding boxes and report the results on the localization set.

4.2 Disease identification

Main results. We compare our model's disease identification performance to four baselines as shown on the left side of Table 1 on the official test split of the ChestXray14 dataset. We did not compare against Guan *et al.* [7] who tested on a different split. The left two [22, 29] are whole image, CAM-based models and the right two [16, 17] are MIL-based models. Note that Li *et al.* [16] and CIA-Net [17] used a combination of 70% unannotated data and 70% annotated cross-validation test data for training. The other models, including ours, are trained only on unannotated data. The results show that our model, which uses region verification and distance learning, outperforms all baselines on average, and is the

Disease	[29]	[22]*	[16]°	[17]°	Ours	w/o RV	w/o DL
Atelectasis	0.700	0.755	<u>0.80</u>	0.79	**0.845**	0.802	0.833
Cardiomegaly	0.810	0.867	<u>0.87</u>	<u>0.87</u>	**0.905**	0.856	0.887
Effusion	0.759	0.815	0.87	**0.88**	<u>0.877</u>	0.849	0.861
Infltration	0.661	0.694	<u>0.70</u>	0.69	**0.817**	0.797	0.799
Mass	0.693	0.802	<u>0.83</u>	0.81	**0.859**	0.826	0.835
Nodule	0.669	0.735	<u>0.75</u>	0.73	**0.824**	0.752	0.792
Pneumonia	0.658	0.698	0.67	<u>0.75</u>	**0.804**	0.751	0.791
Pneumothorax	0.799	0.828	0.87	**0.89**	<u>0.871</u>	0.802	0.858
Consolidation	0.703	0.722	<u>0.80</u>	0.79	**0.810**	0.774	0.803
Edema	0.805	0.835	<u>0.88</u>	**0.91**	0.862	0.813	0.857
Emphysema	0.833	0.856	<u>0.91</u>	**0.93**	0.896	0.757	0.879
Fibrosis	0.786	<u>0.803</u>	0.78	0.80	**0.849**	0.793	0.836
Pleural_Thickening	0.684	0.749	0.79	<u>0.80</u>	**0.829**	0.779	0.805
Hernia	0.872	0.894	0.70	<u>0.92</u>	**0.927**	0.823	0.832
Mean	0.745	0.789	0.81	<u>0.83</u>	**0.855**	0.798	0.833

Table 1: **AUC-ROC score of different disease identification methods.** Our model outperforms state-of-the-art methods, including Wang *et al*. [29], CheXNet [22], Li *et al*. [16], and CIA-Net [17]. The ablation results (gray) of variants of our model without region verification (RV) and distance learning (DL) are better than vanilla DenseNet-121 [22] without these modules. Note that ° denotes using additional bounding box supervision and * denotes based on our implementation; results for other models are from the papers. Bold and underline indicate the best and second best results, respectively.

top-performing model for 10 of the 14 diseases, indicating the effectiveness of joint use of region verification and distance learning for improving disease identification performance.

Our end-to-end framework can be applied to various other backbones with modest hyperparameter tuning such as changing the input image resolution and threshold of CAM. We tested the performance of the vanilla ResNet-50 [8] network, which achieved 0.770 mean AUC-ROC. When ResNet-50 was used as the backbone in our proposed model, it achieved 0.822 mean AUC-ROC, demonstrating the applicability and effectiveness of distance learning and region verification modules when applied to another popular backbone network.

Ablation studies. We show ablation studies in Table 1. We consider the individual contributions of region verification and of distance learning to our model's performance on the right side of Table 1 (in gray). Note that use of both modules improves on the performance of one alone, indicating that they model complementary information. We can further compare the performance of our base model with only distance learning (labeled "w/o RV") or only region verification (labeled "w/o DL") to the column headed [22], which we re-implemented and which serves as a base model to which the region verification and distance learning modules are added. We can note that each module alone improves performance over the baseline.

Learning with hard examples. In our model in Table 1, the hard positive examples included partial label matches (Section 3.2). When positive/negative examples are randomly selected without considering hardness, performance decreases from 0.855 to 0.840, verifying the utility of hard examples. Including partial matches vs. identical matches only performed similarly (0.855 vs. 0.853), indicating positive examples need not include partial matches.

T (IoU)	Method	Ate.	Car.	Eff.	Inf.	Mas.	Nod.	Pn1	Pn2	Mean
0.1	[29]	0.69	0.94	0.66	0.71	0.40	0.14	0.63	0.38	0.57
	[16]$^\diamond$	**0.71**	0.98	**0.87**	**0.92**	**0.71**	**0.40**	0.60	**0.63**	**0.73**
	[17]	0.39	0.90	0.65	0.85	0.69	0.38	0.30	0.39	0.60
	Ours	0.59	**0.99**	0.85	0.76	0.61	0.23	**0.68**	0.49	0.65
0.3	[29]	0.24	0.46	0.30	0.28	0.15	0.04	0.17	0.13	0.22
	[16]$^\diamond$	0.36	0.94	**0.56**	0.66	0.45	**0.17**	0.39	**0.44**	**0.50**
	[17]	0.34	0.71	0.39	0.65	**0.48**	0.09	0.16	0.20	0.38
	Ours	**0.51**	**0.96**	**0.56**	**0.67**	0.45	0.16	**0.43**	0.21	**0.50**
0.5	[29]	0.05	0.18	0.11	0.07	0.01	0.01	0.03	0.03	0.06
	[16]$^\diamond$	0.14	0.84	**0.22**	0.30	0.22	**0.07**	0.17	**0.19**	**0.27**
	[17]	0.19	0.53	0.19	**0.47**	**0.33**	0.03	0.08	0.11	0.24
	Ours	**0.20**	**0.92**	0.19	0.39	0.20	0.06	**0.18**	0.04	**0.27**
0.7	[29]	0.01	0.03	0.02	0.00	0.00	0.00	0.01	0.02	0.01
	[16]$^\diamond$	0.04	0.52	0.07	0.09	0.11	**0.01**	**0.05**	0.05	0.12
	[17]	**0.08**	0.30	**0.09**	**0.25**	**0.19**	0.01	0.04	**0.07**	**0.13**
	Ours	0.04	**0.72**	0.03	0.15	0.02	0.00	0.03	0.01	**0.13**

Table 2: **Accuracy (in %) of different disease localization methods under various T (IoU).** Using the CAM [34] weakly-supervised localization method, our model is on par or even surpasses SOTA methods, including Wang *et al*. [29], Li *et al*. [16], and CIA-Net [17]. $^\diamond$ denotes using additional bounding box supervision. Pn1: Pneumonia. Pn2: Pneumothorax.

Model	IoU (0.1)	IoU (0.3)	IoU (0.5)	IoU (0.7)
Full model	0.65	0.50	0.27	0.13
w/o DL	0.62	0.49	0.27	0.06
w/o RV	0.60	0.48	0.24	0.10

Table 3: **Ablation studies** on distance learning (DL) and region verification (RV) modules for disease localization. We show overall accuracy with different thresholds.

4.3   Disease localization

Main results. We compare the localization performance of our model to the three baseline models used in Table 2 that provided quantitative localization results. For the MIL-based models [16, 17], we compare when the models are trained without any test data; that is, when no disease bounding boxes are used (in cross-validation). The performance of our model was similar to that of the best performing model, Li *et al*. [16], except when T (IoU) was 0.1, where it was second best.

Ablation studies. We analyze the effect of different modules for disease localization. In Table 3, we show ablation results of full models without distance learning and region verification, respectively. Note that each of the modules does contribute to the overall performance of the model, and that overall, region verification contributed more strongly. This stronger contribution may be expected since region verification tends to insure that the identified region contains the features of the targeted disease.

Qualitative results. We show qualitative results in Figure 3 for one example of each of the 8 diseases with location labels. We observe for these images that the probability of the
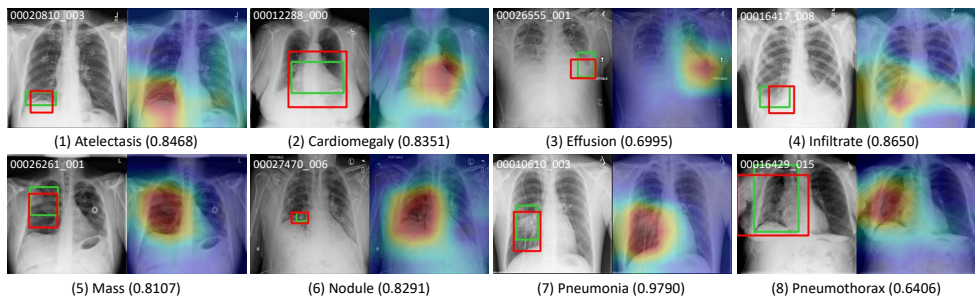
Figure 3: **Qualitative results of 8 disease types.** For each sample, we show the original image, CAM-based [34] heat map, localization result as well as confidential score. Green and red boxes indicate ground truth and predicted result, respectively. Image IDs are in the top-left corner.

disease, shown next to the disease name, indicates that the disease is present and that the high activations in the heatmap are aligned with the ground truth disease locations.

5 Discussion

In this paper we proposed the use of distance learning and region verification for disease identification and localization. We used disease-specific region features which provide a variable size context. Our ablation studies indicate that this strongly improves disease identification and localization performance. This strong contribution may be expected since region verification tends to insure that the identified region contains the features of the targeted disease. We also investigated the use of distance learning with hard example mining for a multi-label task. We observed that use of triplets containing hard examples improves performance, and that the positive examples can be selected from either identical disease label matches only or also include partial label matches. Our experiments showed that each method individually leads to performance improvement, and together offer state-of-the-art performance for disease identification and competitive performance for localization.

References

[1] Vassileios Balntas, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Learning local feature descriptors with triplets and shallow convolutional neural networks. In *BMVC*, 2016.

[2] Yaniv Bar, Idit Diamant, Lior Wolf, Sivan Lieberman, Eli Konen, and Hayit Greenspan. Chest pathology detection using deep learning with non-medical training. In *ISBI*, 2015.

[3] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *arXiv preprint arXiv:1901.07441*, 2019.

[4] Jinzheng Cai, Le Lu, Adam P Harrison, Xiaoshuang Shi, Pingjun Chen, and Lin Yang. Iterative

attention mining for weakly supervised thoracic disease pattern localization in chest x-rays. In *MICCAI*, 2018.

[5] Po-Hsuan Cameron Chen, Yun Liu, and Lily Peng. How to develop machine learning models for healthcare. *Nature Materials*, 18(5):410, 2019.

[6] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[7] Qingji Guan, Yaping Huang, Zhun Zhong, Zhedong Zheng, Liang Zheng, and Yi Yang. Thorax disease classification with attention guided convolutional neural network. *Pattern Recognition Letters*, 131:38–45, 2020.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[9] Xiangteng He, Yuxin Peng, and Junjie Zhao. Fast fine-grained image classification via weakly supervised discriminative localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(5):1394–1407, 2018.

[10] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[11] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, 2017.

[12] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI*, 2019.

[13] Jyun-Yu Jiang, Francine Chen, Yan-Ying Chen, and Wei Wang. Learning to disentangle interleaved conversational threads with a siamese hierarchical network and similarity ranking. In *NAACL*, 2018.

[14] Alistair EW Johnson, Tom J Pollard, Seth Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. MIMIC-CXR: a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 1(2), 2019.

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[16] Zhe Li, Chong Wang, Mei Han, Yuan Xue, Wei Wei, Li-Jia Li, and Li Fei-Fei. Thoracic disease identification and localization with limited supervision. In *CVPR*, 2018.

[17] Jingyu Liu, Gangming Zhao, Yu Fei, Ming Zhang, Yizhou Wang, and Yizhou Yu. Align, attend and locate: Chest x-ray diagnosis via contrast induced attention network with limited supervision. In *ICCV*, 2019.

[18] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. SSD: Single shot multibox detector. In *ECCV*, 2016.

[19] Wu Liu, Cheng Zhang, Huadong Ma, and Shuangqun Li. Learning efficient spatial-temporal gait features with deep learning for human identification. *Neuroinformatics*, 16(3-4):457–471, 2018.

[20] Chih-Yao Ma, Yannis Kalantidis, Ghassan AlRegib, Peter Vajda, Marcus Rohrbach, and Zsolt Kira. Learning to generate grounded image captions without localization supervision. *arXiv preprint arXiv:1906.00283*, 2019.

[21] Yanbo Ma, Qiuhao Zhou, Xuesong Chen, Haihua Lu, and Yong Zhao. Multi-attention network for thoracic disease classification and localization. In *ICASSP*, 2019.

[22] 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.

[23] Pranav Rajpurkar, Jeremy Irvin, Robyn L Ball, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis P Langlotz, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists. *PLoS medicine*, 15(11):e1002686, 2018.

[24] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016.

[25] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[26] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *CVPR*, 2015.

[27] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017.

[28] Ahmed Taha, Yi-Ting Chen, Teruhisa Misu, Abhinav Shrivastava, and Larry Davis. Boosting standard classification architectures through a ranking regularizer. In *WACV*, 2020.

[29] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chest X-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *CVPR*, 2017.

[30] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M Summers. TieNet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *CVPR*, 2018.

[31] Christoph Zauner. Implementation and benchmarking of perceptual image hash functions. *MS thesis*, 2010.

[32] Cheng Zhang, Wu Liu, Huadong Ma, and Huiyuan Fu. Siamese neural network based gait recognition for human identification. In *ICASSP*, 2016.

[33] Wenzhao Zheng, Zhaodong Chen, Jiwen Lu, and Jie Zhou. Hardness-aware deep metric learning. In *CVPR*, 2019.

[34] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016.