Localization with Limited Annotation for Chest X-rays

Eyal Rozenberg   EYALR@CAMPUS.TECHNION.AC.IL
Department of Computer Science
Technion - Israel Institute of Technology. Haifa, Israel

Daniel Freedman   DANIELFREEDMAN@GOOGLE.COM
Google Research
Haifa, Israel

Alex Bronstein   BRON@CS.TECHNION.AC.IL
Department of Computer Science
Technion - Israel Institute of Technology. Haifa, Israel

Abstract

Localization of an object within an image is a common task in medical imaging. Learning to localize or detect objects typically requires the collection of data that has been labelled with bounding boxes or similar annotations, which can be very time consuming and expensive. A technique that could perform such learning with much less annotation would, therefore, be quite valuable. We present such a technique for localization with limited annotation, in which the number of images with bounding boxes can be a small fraction of the total dataset (e.g. less than 1%); all other images only possess a whole image label and no bounding box. We propose a novel loss function for tackling this problem; the loss is a continuous relaxation of a well-defined discrete formulation of weakly supervised learning and is numerically well-posed. Furthermore, we propose a new architecture which accounts for both patch dependence and shift-invariance, through the inclusion of CRF layers and anti-aliasing filters, respectively. We apply our technique to the localization of thoracic diseases in chest X-ray images and demonstrate state-of-the-art localization performance on the ChestX-ray14 dataset introduced in Wang et al. (2017).

1. Introduction

Large-scale labelled datasets are one of the key ingredients in the rapidly developing domain of computer-aided diagnosis systems. In particular, deep learning, which has come to dominate the field of medical imaging in the same way that it has taken over computer vision more generally, is a very data-hungry approach. The combination of large labelled datasets with deep learning techniques has resulted in state-of-the-art (SOTA) algorithms in many medical imaging tasks, including in classification, detection, and segmentation.

A problematic aspect of this approach is the cost of labelling, particularly in localization tasks – either detection or segmentation. In detection, one wishes to find an object in an image by placing a bounding box around it; in the more fine-grained segmentation, one wishes to localize an object with pixel-level granularity. Generally, in order to use deep learning to train networks to perform either of these tasks in standard fashion, one requires a fair amount of annotated images that mirrors the desired output: bounding boxes in the case of detection, and pixel-level masks in the case of segmentation. The major problem
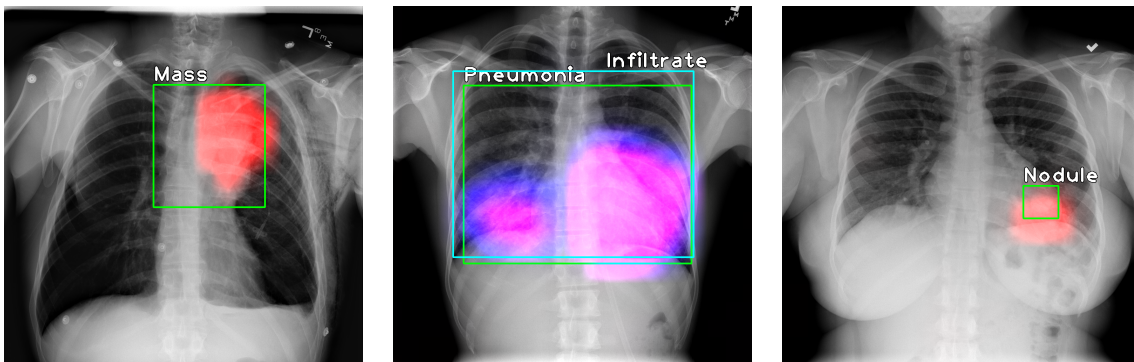
Figure 1: Examples of localized pathologies on test images. The colored blob is the result of our localization algorithms, while the ground truth is marked by a bounding box. In the middle image, two diseases are present, and therefore it is marked twice.

is that collecting such annotations can be very expensive. Indeed, these annotations are much more expensive than their counterparts in the corresponding classification task, in which the labeller must simply specify a label for the image. This problem plagues the general computer vision problems of detection and segmentation but is even worse in the case of medical imaging: the labeller must generally be a physician, resulting in a very costly labelling procedure.

Our goal in this paper is to learn to perform localization with considerably less annotation. In particular, we consider the following setting: relatively cheap whole image (i.e. classification-style) labels are available for each image in the dataset, but only a very small number of examples have a bounding box or segmentation mask labels. This sort of weakly supervised task has been studied before in the computer vision literature, see Babenko (2008). In medical imaging, Li et al. (2018) proposed an approach in the spirit of multiple instance learning and achieved SOTA results on the ChestX-ray14 dataset introduced in Wang et al. (2017).

Despite the success of the approach of Li et al. (2018), it has a number of shortcomings, both in terms of the underlying probabilistic model – which, for example, assumes patch independence, as well in terms of the numerical problems that arise from the formulation. We propose a new technique in which these problems are resolved. In particular, our contributions are twofold: First, we propose a novel loss function, which is a continuous relaxation of a well-defined discrete formulation of weakly supervised learning, and which is numerically well-posed; and secondly, we propose a new architecture which accounts for both patch dependence and shift-invariance, through the inclusion of CRF layers and anti-aliasing filters, respectively. Using the new technique, we show SOTA results for the localization task on the ChestX-ray14 dataset.

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 describes the new technique, focusing alternately on the new loss function and the proposed architecture. Section 4 presents the experimental results, and Section 5 concludes the paper.

2. Related Work

As our results are presented on the ChestX-ray14 dataset of Wang et al. (2017), we begin with a brief discussion thereof. The dataset is a collection of over 100K front-view X-ray images. Using automatic extraction methods from the associated radiological reports by natural language processing, each image was labelled with up to 14 different thoracic pathology classes; in addition, a small number of images with pathology were annotated with hand-labeled bounding boxes for a subset of 8 of the 14 diseases.

Simultaneous with the release of the dataset, Wang et al. (2017) also presented the first benchmark for classification and localization by weakly-supervised CNN architectures pre-trained on ImageNet. This benchmark used only the whole label images for training, ignoring the bounding box annotations. Following the release of the dataset and initial benchmark, several notable works proposed more sophisticated networks for more accurate classification or localization results. Yao et al. (2018) leveraged the inter-dependencies among all 14 diseases using long short-term memory (LSTM) networks without pre-training, outperforming Wang et al. (2017) on 13 of 14 classes. Rajpurkar et al. (2017) proposed classifying multiple thoracic pathologies by transfer-learning with fine-tuning, using a 121-layer Dense Convolutional Network (DenseNet, described in Huang et al. (2016)), yielding SOTA results for the classification task for all 14 diseases. Unlike both previous methods, which do not exploit any of the bounding box annotations, Li et al. (2018) took advantage of these annotations to simultaneously perform disease identification and localization through the same underlying model. Although their method did not attain the top classification results, they did achieve a new SOTA for localization. Subsequent works, such as Guan et al. (2018) and Gündel et al. (2019), have focused on improving the SOTA in classification; by contrast, our focus is localization rather than classification, so we shall not elaborate further on these results.

Another generally related area of interest is object detection. Current SOTA object detectors are of two types: two-stage or single-stage. The two-stage family is represented by the R-CNN framework of Girshick et al. (2014), and is comprised of a region proposal stage followed by the application of a classifier to each of these candidate proposals. This architecture, through a sequence of advances described in Girshick (2015); Ren et al. (2015); He et al. (2017), consistently achieves SOTA results on the challenging COCO benchmark Lin et al. (2014). The initial single-stage detectors, such as YOLO of Redmon et al. (2016) and SSD of Liu et al. (2016), exhibited greater run-time speed at the expense of some accuracy. More recently Lin et al. (2017) proposed RetinaNet, whose training is based on the "focal loss"; this network was able to match the speed of previous single-stage detectors while surpassing the accuracy of all existing SOTA two-stage detectors. These detection approaches, however, are not aimed for tasks that contain a small number of annotated samples as in our setting of interest and are often prone to low accuracy on such datasets.

Finally, we mention the multiple instance learning (MIL) literature, see Babenko (2008). Examples of MIL in medical imaging include Yan et al. (2016); Zhu et al. (2017); Hou et al. (2016).
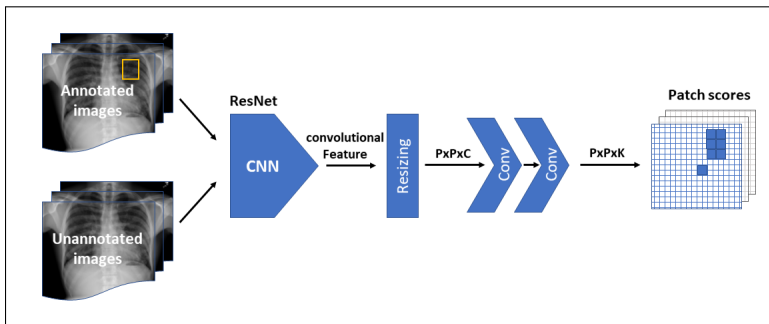
Figure 2: Overview of the base model of Li et al. (2018). Input images are processed a CNN, extracting their feature maps. The latter are then resized and processed by two subsequent convolutional layers to finally output a $P \times P \times K$ tensor of patch scores.

3. Localization with Limited Annotation

3.1. Problem Formulation

Setup: The Base Model As our starting point, we take the approach of Li et al. (2018), which proposes a technique for the classification and localization of abnormalities in radiological images. This approach is very appealing in that it allows for localization to be achieved with a very limited number of bounding box annotations. We now give a brief summary of the technique.

The architecture used in Li et al. (2018) is shown in Figure 2. A preact-ResNet network (see He et al. (2016)), with the final classification layer and global pooling layer removed, is used as the backbone; this part of the architecture encodes the images into a set of feature maps. These feature maps are subsequently divided into a $P \times P$ grid of patches. Through an application of two convolutional layers (including batch normalization and ReLU activation), the number of channels is modified to K, where K is the number of possible disease types. A per-patch probability for each disease class is then derived by the application of a sigmoid function; this is denoted p_j^k, where the probability is that the jth patch of the image belongs to class k. Note that a sigmoid function is applied, rather than a softmax, as a particular patch may belong to more than one disease.

As mentioned above, it is assumed that some images have bounding box annotations, while most do not. Let us define some terms: the image is x; for a disease k, the label $y^k = 1$ if the disease is present, $y^k = 0$ otherwise; if the disease k is annotated with a bounding box b^k, then $a^k = 1$, otherwise $a^k = 0$. Now, the loss function can then be broken into two cases, in terms of whether a bounding box annotation is supplied or not. In the case in which there is a bounding box for a disease of class k, i.e. $a^k = 1$, the bounding box b^k is a subset of the P^2 patches. Then the annotated loss is taken to be

$$L_{ann}^k = -\log p(y^k = 1 | x, b^k)$$

4

where $p(y^k = 1|x, b^k)$ denotes the probability that disease k is within bounding box b^k of image x, and is given by

$$p(y^k = 1|x, b^k) = \prod_{j \in b^k} p_j^k \prod_{j \in \bar{b}^k} (1 - p_j^k) \tag{1}$$

and $\bar{b}^k$ is the complement of bounding box b^k. The above formula is simply the standard formula for combining independent patch probabilities. In the case in which no bounding box is supplied, i.e. $a^k = 0$, the unannotated loss is

$$L_{un}^k = -y^k \log p(y^k = 1|x) - (1 - y^k) \log(1 - p(y^k = 1|x))$$

where

$$p(y^k = 1|x) = 1 - \prod_j (1 - p_j^k) \tag{2}$$

The latter probability is simply the probability that there is at least one patch with disease k, again assuming independence of patches. Finally, the overall loss per image is

$$L = \sum_k \left(\lambda_{ann} a^k L_{ann}^k + (1 - a^k) L_{un}^k \right) \tag{3}$$

We refer to this model – the architecture and the loss – as the *base model*. It was shown to attain SOTA performance in terms of localization on the NIH Chest X-ray dataset introduced in Wang et al. (2017).

Issues with the Approach of Li *et al.* The probabilistic formulation of Li *et al.* is a nice approach to localization tasks with a very limited number of bounding box annotations. In spite of this, there are several issues with the technique that limit its performance:

1. **Single Patch $\Rightarrow$ Positive Declaration:** In Equation (2) of the above derivation for the unannotated loss, only a single patch needs to be positive for a positive disease detection within the image. In general, this assumption is prone to false positives. One would like for multiple patches to be present for a declaration; in particular, a single positive detection could easily be caused by noise. We would, therefore, like to do away with this assumption, by using a novel loss function.

2. **Numerical Issues:** The paper refers to a particular numerical problem that results from the multiplication of many small numbers, as in Equation (1). This numerical underflow is fatal to the approach, and the problem is circumvented in Li et al. (2018) through a series of unjustified heuristics. We propose a formulation of the loss in which these issues never arise.

3. **Patch Independence:** In both Equations (1) and (2), the probabilities of each patch containing an object of a particular class (p_j^k) are treated as independent between patches. This is not correct in practice as we would like to integrate more elaborate terms that model contextual relationships between object/patch classes. We solve this through the use of a Conditional Random Field (CRF) model.

4. **Lack of Shift-Invariance:** As has been pointed out by Zhang (2019), modern CNNs are not technically shift-invariant. In order to improve the performance of the localization, one can, therefore, address this issue through the addition of anti-aliasing filters prior to downsampling, as suggested in Zhang (2019).

Summary of the Proposed Approach. To summarize, our technique is related to the base model but is differentiated in two key ways:

1. The use of a novel loss function, which addresses many of the aforementioned issues.

2. Modifications to the architecture, specifically (a) the incorporation of conditional random field (CRF) layers and (b) the inclusion of anti-aliasing filters.

We now elaborate on each of these, in turn.

3.2. The New Loss Function

Notation As described above, the output of the base model is a tensor of shape the of $P \times P \times K$; we will continue to denote the output as p_j^k, the probability is that the j-th patch of the image belongs to class k. This output will then be fed into a series of layers that implement a CRF model, with p_j^k representing the unary terms in the CRF. The output of the CRF is denoted as z_j^k. We will discuss the details of the CRF model in Section 3.3; for now, we may think of z_j^k as a sharper estimate of whether a particular disease k is present in patch j. z^k indicates the length P^2 vector of all patch values for a given disease k.

The Loss Function: First Pass We first consider an annotated example with $a^k = 1$, i.e. one with a bounding box. As above, the bounding box b^k, consisting of a subset of the patches, contains an object of class k. In this case, the following loss function is natural:

$$L_{ann}^k = -\mathbb{I}[z^k \text{ contains a blob of size} \geq \tau^k |b^k| \text{ within } b^k$$
$$\text{and contains blobs of total size} \leq \rho^k |\bar{b}^k| \text{ within } \bar{b}^k] \tag{4}$$

where $\mathbb{I}[\cdot]$ is the indicator function; $\tau^k, \rho^k \in [0,1]$ are thresholds; and $\bar{b}^k$ is the complement of the bounding box b^k. A blob may be made precise as a connected component; however, this will not lead to a nice differentiable loss. So we make the following continuous relaxation of the above discrete formulation:

$$L_{ann}^k = -\sigma(\mathbf{1}^T(\mathbb{I}[b^k] \odot z^k) - \tau^k |b^k|) \cdot \sigma(\rho^k |\bar{b}^k| - \mathbf{1}^T(\mathbb{I}[\bar{b}^k] \odot z^k)) \tag{5}$$

where $\mathbb{I}[b^k]$ is now an indicator function on the bounding box; $\odot$ is the Hadamard product; $\mathbf{1}$ is the vector of all 1's; and σ is a sigmoid function, i.e. a smooth approximation to the indicator function. What this says is that there must be a total of $\tau^k |b^k|$ patches within b^k which detect class k; the relaxation is that the total no longer has to be in a single connected component. This is a reasonable relaxation, especially since the CRF already encourages smoothness. In addition, we require that there be fewer than $\rho^k |\bar{b}^k|$ patches outside of b^k which detect class k.

Note that this logic extends in a straightforward manner to the case of an unannotated example with $a^k = 0$, when there is no bounding box specified. In the case of a positive patch (i.e. one in which disease k is present), the loss is simply

$$L_{un|pos}^k = -\sigma(\mathbf{1}^T z^k - \hat{\tau}^k) \tag{6}$$

so that the threshold $\hat{\tau}^k$ now has a meaning in absolute terms, i.e. the absolute number of patches vs. the number of patches relative to the size of a bounding box. In the case of a negative patch – where disease k is absent – we have an equation analogous to (6):

$$L_{un|neg}^k = -\sigma(\hat{\rho}^k - \mathbf{1}^T z^k) \tag{7}$$

where $\hat{\rho}^k$ is another threshold, whose meaning is in terms of the absolute number of patches, similar to $\hat{\tau}^k$. Finally, we can combine Equations (6) and (7) to get

$$L_{un}^k = -y^k \sigma(\mathbf{1}^T z^k - \hat{\tau}^k) - (1 - y^k)\sigma(\hat{\rho}^k - \mathbf{1}^T z^k) \tag{8}$$

Addressing Issues (1) and (2). The above formulation addresses Issues (1) and (2) raised in Section 3.1. Regarding Issue (1), Equation (6) requires more than a single patch in order to make a positive declaration; the number of patches must be equal to $\hat{\tau}^k$, which is a per-class parameter which can be chosen. Regarding Issue (2), neither Equations (5) or (8) involve the multiplication of many small values; indeed, both are well-posed from a numerical point of view.

Dealing with Vanishing Gradients Due to the presence of sigmoid functions in Equations (5) and (8), in practice, we experience issues of vanishing gradients during training. We propose the following remedy, based on a different relaxation. We replace Equation (5) with

$$L_{ann}^k = |b^k|^{-1} \text{ReLU}(\tau^k |b^k| - \mathbf{1}^T(\mathbb{I}[b^k] \odot z^k)) + |\bar{b}^k|^{-1} \text{ReLU}(\mathbf{1}^T(\mathbb{I}[\bar{b}^k] \odot z^k) - \rho^k |\bar{b}^k|) \tag{9}$$

Note that there are three fundamental differences between Equations (5) and (9). First, we have replaced the sigmoid functions σ with ReLU functions; this has the effect of still leading to minimal loss (in this case, zero) once the constraints are satisfied, but leads to more nicely behaved gradients. Second, we have replaced the multiplication with addition. Once sigmoids have been replaced by ReLU's, the notion of a "fuzzy and" relaxation is no longer relevant; in this case, addition makes more sense, and again leads to better-behaved gradients. Finally, sigmoids are scaled between 0 and 1, whereas ReLU's can grow without bound; this necessitates the insertion of a scaling factor of $|b^k|^{-1}$ and $|\bar{b}^k|^{-1}$, to ensure that the two terms in the sum are properly balanced.

Similarly, for unannotated samples we replace Equation (8) with:

$$L_{un}^k = y^k \text{ReLU}(\hat{\tau}^k - \mathbf{1}^T z^k) + (1 - y^k) \text{ReLU}(\mathbf{1}^T z^k - \hat{\rho}^k) \tag{10}$$

The thresholds $\tau^k, \hat{\tau}^k, \rho^k, \hat{\rho}^k$ can be treated as parameters of the network, that can be optimized during training, or can be considered hyperparameters. Note that the losses described in Equations (9) and (10) do not suffer from any numerical issues. This is due to the fact that they are not the product of many individual probabilities; rather, they aggregate information across patches in such a way that the resulting loss is numerically stable.

7

Balancing Factors and the Final Loss Function There are two sources of data imbalance to account for. The first is the large imbalance of negative (non-diseased) vs. positive (diseased) examples in the data. To deal with this, we modify Equation (10) slightly, to read

$$L_{un}^k = y^k \operatorname{ReLU}(\hat{\tau}^k - \mathbf{1}^T z^k) + \gamma(1 - y^k) \operatorname{ReLU}(\mathbf{1}^T z^k - \hat{\rho}^k) \tag{11}$$

where γ is the ratio of positive to negative examples in the data. The latter de-emphasizes negative examples relative to the positive examples and follows the practice in Wang et al. (2017).

The second form of imbalance we must account for is that between the annotated and unannotated examples; in practice, there are many more unannotated examples. However, this is already accounted for by the factor λ_{ann} in Equation (3), which is set to a large value. Combining our own annotated and unannotated losses in Equations (9) and (11), respectively, using Equation (3), we arrive at the final form of the per-example loss:

$$
\begin{aligned}
L(x, y, a, b) = \sum_k \Bigg\{ & \lambda_{ann}\, a^k \bigg[|b^k|^{-1} \operatorname{ReLU}(\tau^k |b^k| - \mathbf{1}^T (\mathbb{I}[b^k] \odot z^k(x))) \\
& \qquad\qquad + |\bar{b}^k|^{-1} \operatorname{ReLU}(\mathbf{1}^T (\mathbb{I}[\bar{b}^k] \odot z^k(x)) - \rho^k |\bar{b}^k|) \bigg] \\
& + (1 - a^k) \bigg[y^k \operatorname{ReLU}(\hat{\tau}^k - \mathbf{1}^T z^k(x)) + \gamma(1 - y^k) \operatorname{ReLU}(\mathbf{1}^T z^k(x) - \hat{\rho}^k) \bigg] \Bigg\}
\end{aligned}
\tag{12}
$$

where the dependence of the z variables on the image x has been made explicit.

3.3. Architectural Modifications

CRF Model As mentioned in Issue (3) in Section 3.1, we would like to do away with the assumption of patch independence; and indeed, in our derivation of Equation (12), we did not make use of such assumptions. However, to further bolster the dependence between neighboring patches, we introduce a CRF model into our network. The CRF introduces, in an explicit manner, a spatial dependency between patches. The effect of the CRF is to increase the confidence for a given patch's predicted label, and thereby to improve localization. There are several choices amongst neural network-compatible CRFs; we choose the recent pixel-adaptive convolution (PAC) approach of Su et al. (2019), due to its simplicity and excellent performance. We thus integrate the PAC-CRF modification to our base network and train the model end-to-end.

Given the patch probability outputs p_j^k of the base model, the unary potentials of the CRF are simply taken as the $\psi_j^{(u)}(k) = p_j^k$. Thus, in the absence of neighbor dependence, the CRF will simply choose $z_j^k = p_j^k$. Neighbour dependence is introduced through pairwise potentials. As in Su et al. (2019), we take this potential to be $\psi_{jl}^{(p)}(k_j, k_l) = G(f_j, f_l) W_{k_j k_l}(\xi_j - \xi_l)$, where f are a set of learnable features on the $P \times P$ grid; G is a fixed Gaussian kernel; ξ_j is the pixel coordinates of patch j; and W is the inter-class compatibility function, which varies across different spatial offsets, and is also learned. The pairwise connections are defined over a fixed window Ω around each patch.

Our training regime for the CRF is as follows. First, we train the unary terms using the base model until converges. Then, we freeze the base model and train only the PAC-CRF
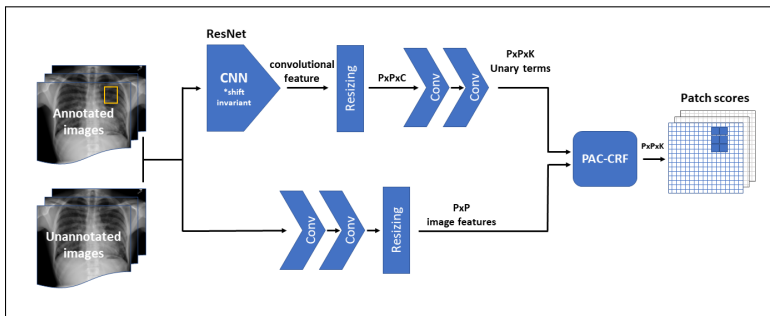
Figure 3: Our model consists of two branches: the upper one has the same architecture as in Figure 2, though modified with anti-aliasing filters, and computes the unary terms of the CRF model. The lower branch extracts features from the input images to form a feature-tensor of the same size as the unary terms, $P \times P$, which are used in the pairwise terms. Both enter into the PAC-CRF of Su et al. (2019), outputting a $P \times P \times K$ tensor of patch scores.

part, using a 19×19 PAC filter. As our unary model outputs a $P \times P \times K$ tensor, we add to our PAC-CRF model two 2D convolution layers, each followed by a rectified linear unit and batch-normalization, to output a tensor with the same size.

Anti-Aliasing An important property of any model whose goal is to perform localization or segmentation is that the output of the model should be shift-invariant with regard to its input. However, as Zhang (2019) has noted, standard CNNs use downsampling layers while ignoring sampling theorem, and are therefore not shift-invariant; this is in spite of the fact that CNNs are commonly used as the backbone of many localization/segmentation tasks. To circumvent this problem, an anti-aliasing filter is required prior to every downsampling part of the network. In particular, Zhang (2019) proposed the insertion of a blur kernel as a low-pass filter prior to each downsampling step in the network; and thereby demonstrated an increased accuracy across several commonly used architectures and tasks. Following Zhang (2019), we thus modify the backbone of our base model and integrate such low-pass filters as part of the preact-ResNet network of He et al. (2016). This effectively addresses Issue (4) in Section 3.1.

4. Results

Dataset As in Li et al. (2018), We have examined our model over NIH Chest X-ray dataset (Wang et al. (2017)). The NIH Chest X-ray dataset consists of 112,120 frontal-view X-ray images with 14 disease labels (each image can have multi-labels); images can have no label as well. Out of the more than 100K images, the dataset contains only 880 images with bounding box annotations; some images have more than one such box, so there are a total of 984 labelled bounding boxes. The remaining 111,240 images have no bounding box annotations but do possess class labels. The images are 1024×1024, but we have resized them 512×512 for faster processing; we have also normalized the image range to $[-1, 1]$.

The 984 bounding boxes annotations are only given for 8 of the 14 disease types. Despite this fact, we have noticed that we get superior results by continuing to learn on the full

Table 1: IoU and IoR disease localization accuracy, using 80% of annotated samples and 0% or 20% of unannotated samples (selected randomly) for training. Evaluation set composed of the remaining 20% annotated-samples of each fold. Showing significant improvement over SOTA results.

IoU accuracy									
$un[\%]$	model	Atelectasis	Cardiomegaly	Effusion	Infiltration	Mass	Nodule	Pneumonia	Pneumothorax
0%	ours	**0.818**	**1**	**0.882**	**0.927**	**0.695**	**0.404**	**0.918**	**0.726**
	ref.	0.488	0.989	0.693	0.842	0.342	0.081	0.715	0.437
20%	ours	**0.779**	**1**	**0.843**	**0.945**	**0.709**	**0.444**	**0.915**	**0.732**
	ref.	0.687	0.978	0.831	0.9	0.634	0.241	0.568	0.576
IoR accuracy									
$un[\%]$	model	Atelectasis	Cardiomegaly	Effusion	Infiltration	Mass	Nodule	Pneumonia	Pneumothorax
0%	ours	**0.889**	**1**	**0.92**	**0.95**	**0.773**	**0.58**	**0.933**	**0.767**
	ref.	0.528	1	0.753	0.875	0.452	0.111	0.786	0.473
20%	ours	**0.844**	**1**	**0.896**	**0.967**	**0.808**	**0.52**	**0.935**	**0.806**
	ref.	0.724	0.991	0.874	0.921	0.674	0.271	0.644	0.624

complement of 14 disease types, which seems to imply some interesting interdependence amongst disease classes.

Evaluation Metrics Although we train over both annotated and unannotated examples, we are interested in localization accuracy, so we evaluate solely over annotated samples. We use Intersection-over-Union (IoU) and Intersection-over-Region (IoR) to measure localization accuracy. A patch is taken to be positive, i.e. the disease k is present in patch j, if its value is greater than 0.5: $z_j^k \geq 0.5$. The union of all positive patches is the detected region. The IoU and IoR can then be computed between the ground truth bounding box and the detected region. A localization is taken to be correct if IoU $\geq T$ or IoR $\geq T$ for given threshold T; following the practice of Li et al. (2018), we use $T = 0.1$. Performance statistics are then computed over 5-fold cross-validation of the annotated samples, as is also done in Li et al. (2018).

Training Details We use ResNet-50 as the backbone of our model. We initialize the weights based on ImageNet (Deng et al. (2009)) pre-training and then allow them to evolve as training proceeds. We take $P = 20$ for a 20×20 patch grid, and take $\lambda_{ann} = 70$. We use the ADAM optimizer, see Kingma and Ba (2014), with a weight decay regularization coefficient equal to 0.01 and exponentially-decaying learning-rate, initialized to 0.001. We take the batch size to be 48.

It is natural to treat the thresholds $\tau^k, \hat{\tau}^k, \rho^k, \hat{\rho}^k$ as parameters of the network, which can be optimized during training. Unfortunately, this quickly leads to the degenerate solution, i.e: $\tau, \hat{\tau} \to 0$, $\rho \to 1$ and $\hat{\rho} \to P^2$. In order to avoid reaching the degenerate solution, we employ the following procedure. We begin by freezing the thresholds, and training only the network weights; we then freeze the network weights, and find the optimal thresholds; we continue to alternate this procedure until convergence.

Experiments We compare our results with those of Li *et al.* which represent the SOTA for the localization task on the NIH Chest X-ray dataset. As has been mentioned above, we use 5-fold cross-validation. We examine two separate settings: (a) the model is trained

using only 80% of the annotated samples, with the 80% representing the training part of the fold; (b) the model is trained using 80% of the annotated samples as described in (a), as well as 20% of the unannotated samples, representing about 20K samples (selected randomly). In both cases, the results are evaluated on the remaining 20% of the annotated samples of each fold.

In Table 1 we present two versions of the localization accuracy, with the top based on IoU and the bottom based on IoR; best results are shown in bold. Our method outperforms that of Li *et al.* for all 8 disease classes, for IoU and IoR, and for both settings - with no extra unannotated data added, and with 20% unannotated data. Examining the IoU data more closely, we see several patterns. First, we perform considerably better than Li et al. (2018) when no unannotated data is added; for example, the accuracy on Atelectasis and Mass is nearly *double* that of Li et al. (2018), whereas the performance on Nodule is *five times* better. Second, with the addition of unannotated data, the gaps narrows – for example, Nodule is now slightly less than double, and many other disease classes have quite a bit smaller gap – but the gap is still present for each disease class. We hypothesize that our improvement is less in most cases simply because the algorithm trained with no unannotated data already has a fairly high performance in most cases; thus, the marginal benefit of adding the unannotated data is smaller. Third, in examining our own results, we see that the addition of unannotated data often helps, but does not always do so. In particular, there is an increase in localization accuracy for four of the eight diseases – Infiltration, Mass, Nodule, Pneumothorax – while two diseases, Cardiomegaly and Pneumonia, undergo little or no change. The remaining two, Atelectasis and Effusion, actually suffer a decrease in accuracy due to the addition of the extra unannotated data. In examining the data, the solution to this puzzle becomes apparent: Atelectasis and Effusion have the largest number of annotated examples of the eight disease classes. This explains why they have quite high localization accuracies to begin with, when no unannotated data has been added (0.818 and 0.882, respectively); and why the addition of extra unannotated examples does not help. (It is interesting to note that Pneumonia has high accuracy and a decent number of annotated examples; the reason it differs from Atelectasis and Effusion is that it also has a relatively small number of *unannotated* examples, so that adding them does not affect its performance too much.)

We note that the IoR data is fairly similar to the IoU data, and most of the observations above hold in this case as well. Qualitative examples of the localizations derived are shown in Figure 1.

5. Conclusions

We have presented a new technique for localization with limited annotation. Our method is based on a novel loss function, which is mathematically and numerically well-posed; and an architecture that explicitly accounts for patch non-independence and shift-invariance. We present SOTA results for localization on the ChestX-ray14 dataset. Future work will focus on applying these ideas to the realm of semantic segmentation.

References

Boris Babenko. Multiple instance learning: algorithms and applications, 2008.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

Qingji Guan, Yaping Huang, Zhun Zhong, Zhedong Zheng, Liang Zheng, and Yi Yang. Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification. *ArXiv*, abs/1801.09927, 2018.

Sebastian Gündel, Sasa Grbic, Bogdan Georgescu, Siqi Liu, Andreas Maier, and Dorin Comaniciu. Learning to recognize abnormalities in chest x-rays with location-aware dense networks. In Ruben Vera-Rodriguez, Julian Fierrez, and Aythami Morales, editors, *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pages 757–765, Cham, 2019. Springer International Publishing. ISBN 978-3-030-13469-3.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 630–645, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46493-0.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

Le Hou, Dimitris Samaras, Tahsin M Kurc, Yi Gao, James E Davis, and Joel H Saltz. Patch-based convolutional neural network for whole slide tissue image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2424–2433, 2016.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2016.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, 2017.

Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 21–37, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46448-0.

Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Yi Ding, Aarti Bagul, Curtis P. Langlotz, Katie S. Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *ArXiv*, abs/1711.05225, 2017.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

Hang Su, Varun Jampani, Deqing Sun, Orazio Gallo, Erik Learned-Miller, and Jan Kautz. Pixel-adaptive convolutional neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

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 *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3462–3471, July 2017. doi: 10.1109/CVPR.2017.369.

Zhennan Yan, Yiqiang Zhan, Zhigang Peng, Shu Liao, Yoshihisa Shinagawa, Shaoting Zhang, Dimitris N Metaxas, and Xiang Sean Zhou. Multi-instance deep learning: Discover discriminative local anatomies for bodypart recognition. *IEEE transactions on medical imaging*, 35(5):1332–1343, 2016.

Li Yao, Eric Poblenz, Dmitry Dagunts, Ben Covington, Devon Bernard, and Kevin Lyman. Learning to diagnose from scratch by exploiting dependencies among labels. *ArXiv*, abs/1710.10501, 2018.

Richard Zhang. Making convolutional networks shift-invariant again. *ArXiv*, abs/1904.11486, 2019.

Wentao Zhu, Qi Lou, Yeeleng Scott Vang, and Xiaohui Xie. Deep multi-instance networks with sparse label assignment for whole mammogram classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 603–611. Springer, 2017.