Supplementary material

Scale patch-level probability to avoid numerical underflow

Notation p_{ij}^k and $1-p_{ij}^k$ represent a patch's (the jth patch of image i) positive and negative probabilities for class k. Their values are always in $[0, 1]$. We consider the problem of numerical underflow as follows. The product terms ($\prod$) in Eq.1 and Eq.2 can quickly go to 0 when many of the terms in the product is small due to the limited precision of float numbers. The log loss in Eq. 3 mitigates this for Eq. 1, but does not help Eq. 2, since the log function can not directly affect its product term. This effectively renders Eq. 2 as a constant value of 1, making it irrelevant on updating the network parameters. (The contribution of the gradient from Eq. 2 will be close to 0.) Similar things happen at test time. To do binary classification for an image, we determine its label by thresholding the image-level score (Eq. 2). It is impossible to find a threshold in $[0, 1]$ to distinguish the image-level scores when the score (Eq. 2) is a constant of 1; all the images will be labeled the same.

Fortunately, if we can make sure that the image-level scores $p(y_k|x_i, \mathrm{bbox}_i^k)$'s and $p(y_k|x_i)$ spread out in $[0, 1]$ instead of congregating at 1, we then can find an appropriate threshold for the binary classification. To this end, we normalize p_{ij}^k and $1-p_{ij}^k$ from $[0, 1]$ to $[0.98, 1]$. The reason of such choice is as follows. In the actual system, we often use single-precision floating-point number to represent real numbers. It can represent a real number as accurate as 7 decimal digits [1]. If the number of patches in an image, $m = 16 \times 16$, a real number $p \in [0, 1]$ should be larger than around 0.94 (by obtaining p from $p^{256} \geq 10^{-7}$) to make sure that the p^m varies smoothly in $[0, 1]$ w.r.t. p changes in $[0.94, 1]$. To be a bit more conservative, we set 0.98 as our lower limit in our experiment. This method enables valid and efficient training and testing of our method. And in the evaluation, the number of thresholds can be finite to calculate the AUC scores, as the image-level probability score is well represented using the values in $[0, 1]$. A downside of our approach is that a normalized patch-level probability score does not necessarily reflect the meaning of probability anymore.

Disease Localization Results

Similarly, we investigate the importance of bounding box supervision by using all the unannotated images and increasing the amount of annotated images from 0% to 80% by the step of 20% (Figure 1). without annotated images (the most left bar in each group), the model is only supervised by image-level labels and optimized using probabilistic approximation from patch-level predictions. The results by unannotated images only are not able to generate accurate localization of disease. As we increase the amount of annotated images gradually from 0% to 80% by the step of 20% (from left to right in each group), the localization accuracy for each type is increased accordingly.

Next, we fix the amount of annotated images to 80% and increase the amount of unannotated images from 0% to 100% by the step of 20% to observe whether unannotated images are able to help annotated images to improve the performance (Figure 2). For some diseases, it achieves the best accuracy without any unannotated images. For most diseases, the accuracy experience an accuracy increase, a peak score, and then an accuracy fall (from orange to green bar in each group) as we increase the amount of unannotated images. A possible explanation is that too many unannotated images overwhelm the strong supervision from the small set of annotated images. A possible remedy is to lower the weight of unannotated images during training.

Lastly, We use 80% annotated images and 50% unannotated images to train the model and evaluate on the other 20% annotated images in each fold. Comparing with the reference model [2], our model achieves higher localization accuracy for various T(IoR) as shown in Table 1.

References

[1] I. S. Committee et al. 754-2008 ieee standard for floating-point arithmetic. *IEEE Computer Society Std*, 2008, 2008.

[2] 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. IEEE, 2017.

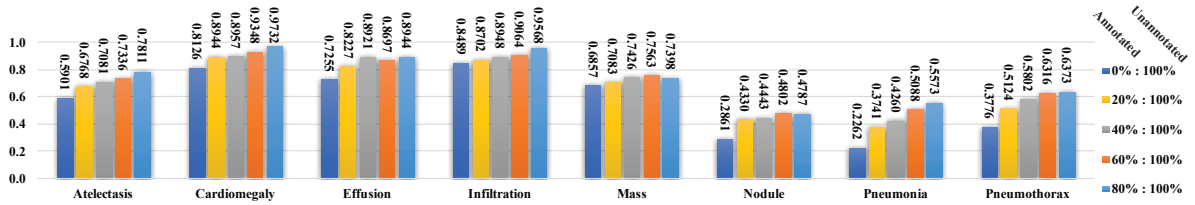

Figure 1. Disease localization accuracy using IoU where T(IoU)=0.1. Training set: annotated samples, {0% (0), 20% (176), 40% (352), 60% (528), 80% (704)} from left to right for each disease type; unannotated samples, 100% (111, 240 images). The evaluation set is 20% annotated samples which are not included in the training set. For each disease, the accuracy is increased from left to right, as we increase the amount of annotated samples, because more annotated samples bring more bounding box supervision to the joint model.

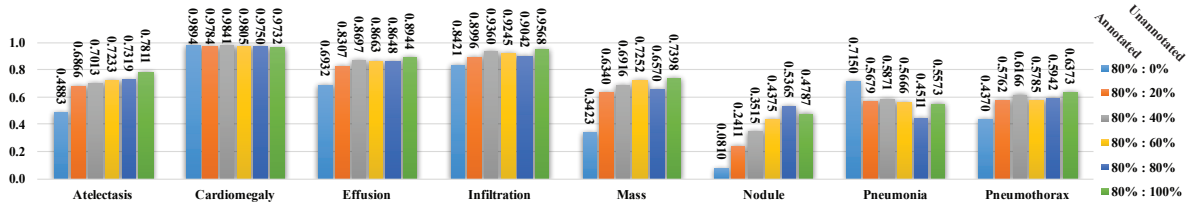

Figure 2. Disease localization accuracy using IoU where T(IoU)=0.1. Training set: annotated samples, 80% (704 images); unannotated samples, {0% (0), 20% (22, 248), 40% (44, 496), 60% (66, 744), 80% (88, 892), 100% (111, 240)} from left to right for each disease type. The evaluation set is 20% annotated samples which are not included in the training set. Using annotated samples only can produce a model which localizes some diseases. As the amount of unannotated samples increases in the training set, the localization accuracy is improved and all diseases can be localized. The joint formulation for both types of samples enables unannotated samples to improve the performance with weak supervision.

T(IoR)	Model	Atelectasis	Cardiomegaly	Effusion	Infiltration	Mass	Nodule	Pneumonia	Pneumothorax
0.1	ref.	0.62	**1.00**	0.80	0.91	0.59	0.15	**0.86**	0.52
	ours	**0.77** $\pm$ 0.06	0.99 $\pm$ 0.01	**0.91** $\pm$ 0.04	**0.95** $\pm$ 0.05	**0.75** $\pm$ 0.08	**0.40** $\pm$ 0.11	0.69 $\pm$ 0.09	**0.68** $\pm$ 0.10
0.25	ref.	0.39	0.99	0.63	0.80	0.46	0.05	**0.71**	0.34
	ours	**0.57** $\pm$ 0.09	**0.99** $\pm$ 0.01	**0.79** $\pm$ 0.02	**0.88** $\pm$ 0.06	**0.57** $\pm$ 0.07	**0.25** $\pm$ 0.10	0.62 $\pm$ 0.05	**0.61** $\pm$ 0.07
0.5	ref.	0.19	0.95	0.42	**0.65**	0.31	0.00	0.48	0.27
	ours	**0.35** $\pm$ 0.04	**0.98** $\pm$ 0.02	**0.52** $\pm$ 0.03	0.62 $\pm$ 0.08	**0.40** $\pm$ 0.06	**0.11** $\pm$ 0.04	**0.49** $\pm$ 0.08	**0.43** $\pm$ 0.10
0.75	ref.	0.09	0.82	0.23	0.44	0.16	0.00	0.29	0.17
	ours	**0.20** $\pm$ 0.04	**0.87** $\pm$ 0.05	**0.34** $\pm$ 0.06	**0.46** $\pm$ 0.07	**0.29** $\pm$ 0.06	**0.07** $\pm$ 0.04	**0.43** $\pm$ 0.06	**0.30** $\pm$ 0.07
0.9	ref.	0.07	**0.65**	0.14	**0.36**	0.09	0.00	0.23	0.12
	ours	**0.15** $\pm$ 0.03	0.59 $\pm$ 0.04	**0.23** $\pm$ 0.05	0.32 $\pm$ 0.07	**0.22** $\pm$ 0.05	**0.06** $\pm$ 0.03	**0.34** $\pm$ 0.04	**0.22** $\pm$ 0.05

Table 1. Disease localization accuracy comparison using IoR where T(IoR)={0.1, 0.25, 0.5, 0.75, 0.9}. The bold values denote the best results. Note that we round the results to two decimal digits for table readability. Using different thresholds, our model outperforms the reference baseline in most cases and remains capability of localizing diseases when the threshold is big. The results for the reference baseline are obtained from the latest update of [2].