CPSC 93 Final Report: Thoracic Disease Identification and Localization

William Colgan WCOLGAN1@SWARTHMORE.EDU
Swarthmore College, 500 College Ave., Swarthmore, PA 19081 USA

Abstract

I train a 121-layer deep convolutional neural network to accurately identify and localize thoracic diseases in chest X-ray images. This is enabled by many skip connections in the network and a new chest X-ray dataset that contains over 100,000 images. I experiment with a number of different optimizers, learning rate decay schedules, and data augmentation strategies. I also increase the resolution of the model and perform limited supervision with bounding boxes to improve the localization. I report near state of the art identification results and good localization results. Localization is essential for models to work with radiologist to diagnose chest X-ray images.

1. Introduction

Computer aided diagnosis has the potential to increase the efficiency of medical practice and decrease medical error. One application is the analysis of medical images, such as X-rays and MRIs. This is enabled by recent advances in machine learning, such as deep convolutional neural networks (DCNNs). Recent successes include the diagnosis of breast cancer and Alzheimers. Some of these models have achieved human level performance, but they are unlikely to replace radiologist for liability reasons. To be deployed these models must work with radiologist. To achieve this, they must provide evidence for the prediction, such as the spatial location of the disease. This has the potential to increase the efficiency of radiologists and prevent them from missing secondary findings like nodules.

In 2017 the NIH released a new dataset with 112,120 frontal-view chest X-ray images (Wang et al., 2017). This dataset is an order a magnitude larger than the previous chest X-ray dataset which enables the training of DCNNs. There have been several papers on the chestX-ray. Rajpurkar et al. reported radiologist level performance in detecting pneumonia (Rajpurkar et al., 2017). This exciting result highlights the potential of this new dataset. However, there have been few papers on predicting the localization of diseases within the chestX-ray images. This this important for the deployment of these models.

I aim to predict both the disease and its localization with the chestX-ray dataset. To do this I use a 121-layer DenseNet, which is the same model that Rajpurkar et al. used. I first compare this model to ResNet-50 and experiment with a number of different optimizers, learning rate decay schedules, and data augmentation strategies. I then use class activation mapping to predict the localization of diseases. To improve the localization, I train DenseNet with limited supervision from bounding boxes and increase the resolution of the prediction. I achieve identification near the current state of the art and good localization. This provides a framework for the development of more complex models.

2. Related Work

There have been several papers on the chestX-ray dataset since it was published in 2017. These primarily focus on supervised disease identification.

Wang et al. published baselines for identification and localization in conjunction with the chestX-ray dataset (Wang et al., 2017). To do this they applied several standard DCNNs. They adapted the networks for multilabel classification by replacing the final softmax layer with a sigmoid layer. They pretrained the networks using ImageNet and experimented with several pooling strategies and loss functions. ResNet-50 consistently outperformed the other networks, max and average pooling were equivalent, and weighted cross entropy was the most effective loss function. Their AUROC values ranged from .564 for mass to .814 for cardiomegaly. To localize the diseases, they used class activation mapping. To generate bounding boxes, they normalized the heatmap, thresholded it, and then drew boxes around the regions. Their IoU values were generally low, which the exception of cardiomegaly.

CPSC 93 Final Report Proceedings, Swarthmore College, Spring 2018.

Yao et al. modeled the statistical dependencies between labels to improve performance (Yao et al., 2017). The idea is that certain findings will increase or decrease the probability of other findings. To do this they used an LSTM to decode the output of the DCNN instead of a logits layer. They used DenseNet as their DCNN because the skip connections make it easier to train. Additionally, they decreased the number of ConvBlocks in each DenseBlock from 16 to 4. This allows them to train their model from scratch with adam and significant data augmentation. This model outperformed the baseline by .05 to .1 AUROC for every disease. However, it is unclear how much of that increase is due to the LSTM since they did not report results without it. They did not predict disease localization.

Rajpurkar et al. also used DenseNet (Rajpurkar et al., 2017). They initially focused on predicting pneumonia because this is particularly challenging for radiologists. They used a standard 121-layer DenseNet which they pretrain using ImageNet. They maximized a weighted cross entropy loss for binary prediction. Using adam and horizontal flipping of images they were able to exceed average radiologist performance on the F1 matric for pneumonia. This evaluation is independent of chest X-ray which are only about 90% accurate. The average radiologist labels and the chestX-ray differed significantly, which is concerning. They then applied the approach to the other 14 diseases in the chest X-ray dataset. They achieved AUROC values between .735 and .923 which are the best values reported to date. They generated class activation maps, but they did not quantify disease localization.

Li et al. used limited bounding box supervision to improve both identification and localization of diseases (Li et al., 2017). To do this they used a convolutional recognition network to decode the output of the DCNN instead of a logits layer. They used ResNet-50 pretrained on ImageNet without pooling. To decrease the size of the output they did patch slicing with max pooling. They then fed this though the recognition network which contained a 3x3 convolution layer, a batch norm layer, and a logits layer. This yielded a disease prediction for each patch which they aggregated to generate an image level prediction. For images with bounding boxes, they calculated the loss for each patch in addition to the image level loss. They got similar AUROC values to Yao et al, and significantly outperformed the baseline for IoU. The greatest improvements were in difficult diseases but these IoU values remain low. For example, they increased the accuracy at the .5 threshold from .001 to .007 for nodule.

3. Methods

Given chest X-ray images with limited bounding box information, I aim to design a model which can accurately identify diseases and their localization in the images

3.1. Architecture

DCNN. I used two standard deep convolutional neural networks, ResNet-50 and DenseNet-121 (He et al., 2015; Huang et al., 2016). Both these networks make use of skip connections to enable faster training, but DenseNet has significantly more skip connections. Within a DenseBlock every layer is connected to every other layer which enables training of networks with greater than 100 layers (figure 1). ResNet-50 is used as a baseline.

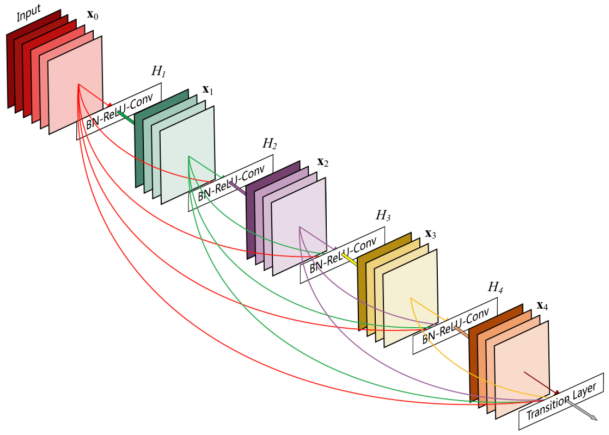


Figure 1. A 5-layer DenseBlock. Each layer takes all preceding feature-maps as input.

Multilabel setup. I formulate the task as a multilabel classification problem. The disease labels for each image are represented as a 14-dimensional vector where each index corresponds to the presence of a disease. An image with no findings is represented as vector with all zeros. To enable multilabel classification, the final softmax layer in the DCNN is replaced by a sigmoid layer. This means that each disease is treated as a binary classification problem.

Localization. To localize diseases in the images I used class activation mapping (CAM) (Zhou et al., 2015). This is a state of the art technique for unsupervised object localization. I did this by applying the logits layer to the DCNN output before global average pooling. I then resized this to 224 x 224 with bilinear interpolation. To generate bounding boxes for each class from the CAM, I then thresholded the heatmap at the value 0.9.

Increasing resolution. To increase the resolution of the bounding box prediction I modified DenseNet-121. I did this by removing all the 2 x 2 average pooling layers. This increased the resolution of the DCNN output from 7 x 7 to 56 x 56 without increasing the number of parameters in the

model.

Cross entropy loss. Initially I used cross entropy loss, but the model struggled to learning positive instances, particularly for rare diseases. The remedy this, I switched to a weighted cross entropy loss like Wang et al (Wang et al., 2017). This uses a positive/negative balancing factor β_P, β_N to ensure that equal weight it positive and negative examples in a given batch. Weighted cross entropy loss (W-CEL) is defines as,

$$W{-}CEL = \beta_P \sum_{y_c=1} -ln(f(x_c)) + \beta_N \sum_{y_c=0} -ln(1-f(x_c)) \quad (1)$$

where β_P is set to $\frac{|P|+|N|}{|P|}$ and β_N is set to $\frac{|P|+|N|}{|N|}$. $|P|$ and $|N|$ are the total number of 1s and 0s in the batch.

Localization loss. For limited supervision with the bounding boxes I used cross entropy loss between the CAMs and bounding boxes. This was calculated only for the bounding boxes or for the bounding boxes and CAMs thresholded at the value 0.9.

3.2. Training

DCNN. Both ResNet-50 and DenseNet-121 are implemented in TensorFlow using the TensorFlow-Slim framework. They are initialized with weights learned on the ImageNet dataset. Training is performed on a single Nvidia GPU, either a 1080 or P5000. The P5000 was needed to fit DenseNet-121 without average pooling.

Data augmentation. Two data augmentation strategies were employed, horizontal flipping and horizontal flipping with random cropping. The random crops ranged in size from 100

Learning rate decay. Two learning rate decay strategies were employed, gradual and step decay. Smooth decay decreased the learning rate by a factor of .94 every 2 epochs. Step decay decreased the learning rate by a factor of .1 every 10 epochs. The initial learning rate for both was .001.

Optimization. Several optimization strategies were experimented with, but the most effective was Adam with the standard parameters ($\beta 1 = .9$ and $\beta 2 = .999$). A batch size of 16 was used except for DenseNet-121 without average pooling. This architecture need a batch size of 8 to fit on a single GPU.

4. Data

I used the chestXray-14 dataset which contains 112,120 frontal-view X-ray images of 30,805 unique patients (Wang et al., 2017). Each image is labeled for 14 common thoracic diseases. The labels are extracted from radiology reports and are estimated to be at least 90% accurate. The dataset also includes 940 labeled bounding boxes for 880 images which were generated by certified radiologists. These are evenly split across 8 of the diseases in the dataset. I resized the 3-channel image from 1024 x 1024 to 224 x 224 for faster processing and normalized the pixel values for each channel to have the same mean and standard deviation as the images in ImageNet.

Unsupervised localization. For the unsupervised localization experiments I used the standard 70% train, 10% validation, and 20% test split for the chestX-ray dataset. There is no patient overlap between the splits. All the bounding boxes are in the test split.

Semi-supervised localization. For the semi-supervised localization experiments I split the chestX-ray dataset into 5 folds with 80% train and 20% test. This was done with stratified random sampling so that there is the same number of bounding boxes for each disease in each split. There is no patient overlap between the splits.

5. Experiments and Results

5.1. Disease Identification

ResNet. I first established a baseline with ResNet to validate my code. This revealed a number of issues with my conversion of the dataset into TFRecord format. Once these were corrected I experimented with different loss functions and training techniques. I was able to achieve similar results to Wang et al., by just tuning the logits layer of a pretrained ResNet (figure 2) (Wang et al., 2017). Using weighted cross entropy loss like Wang et al., significantly increased the AUROC for less common diseases (figure 2). Fully tuning the pretrained ResNet slightly decreased the AUROC. This suggests that decent results can be achieved by using ResNet as a feature extractor and that is is difficult to fully train ResNet with the chestX-ray dataset.

DenseNet. To make training easier I used DenseNet. I experimented with a number of different optimizers, learning rate decay schedules, and data augmentation strategies. Experiments were run for 300,000 steps (75 epochs) to ensure convergence. For horizontal flipping and random cropping, adam was best with the gradual decay, and momentum was best with step decay. These had similar W-CEL and average AUROC values (figure 3). For horizontal flipping, step decay significantly outperformed gradual decay. This is likely because step decay prevents overfitting with less data augmentation. The W-CEL for the gradual decay is very low but this does not correspond to high average AUROC on the validation set (figure 3). Interesting, momentum converged faster than adam with step decay. My AUROC values are slightly worse than Rajpurkar et al (Rajpurkar et al., 2017). This is probably because I am doing multilabel classification instead of binary classification

(figure 4).

5.2. Disease Localization

CAMs. To test the ability of DenseNet to localize diseases without bounding box supervision, I generated class activation maps (CAMs) (Zhou et al., 2015). These heatmaps were remarkably accurate for some images (figure 5). In these images, DenseNet focuses on small regions for localized diseases such as infiltration and nodule, and on larger regions for less localized diseases such as atelectasis and pleural thickening. I cannot truly validate these localizations since I am not a radiolist. For other images the heatmaps are diffuse and non-specific. In many of the images DenseNet focuses on medical devices and lines in the image. These are predictive of sick patients who are more likely to have a disease. This suggests that bounding box supervision is required for the localization of the disease instead of just regions which are predictive.

Bounding box locations. Ideally, supervising with bounding boxes will improve identification as well as localization because it prevents overfitting to areas not associated with the disease and because the spatial location is indicative of the disease. To learn more about the spatial distribution of the bounding boxes I generated a heatmap for each disease with all the bounding boxes (figure 6). There are large differences in spatial deviation and size between diseases. For example, the bounding boxes for cardiomegaly are large and centered on the heart, while the bounding boxes for nodule are small and distributed throughout the lungs. It is clear that the localization is indicative for some diseases. For example, atelectasis is usually in the lower lungs and pneumothorax is usually in the upper lungs.

Semi-supervised localization. To enable limited supervision by the bounding boxes I added localization loss to the total loss calculation. I experimented with calculating the localization loss only for the bounding boxes or for the bounding boxes and CAMs thresholded at the value 0.9. The point of including the thresholded CAMs was to stabilize the loss function, but this did not work. The model minimized the localization loss by predicting the disease was in the top left corner which did not negatively affect the identification loss. Just including the bounding boxes worked better. It significantly increased the IoU for most of the diseases and slightly improved the AUROC (figures 4,7). Some of the localization error appears to be caused by the low resolution (7 x 7) of the prediction. This is especially true for disease with small bounding boxes, like mass (figure 7).

Increasing resolution. To increase the resolution of the prediction I removed the 2 x 2 average pooling layers from DenseNet. This increased the time needed to train the model by a factor of 4. I trained this model with limited

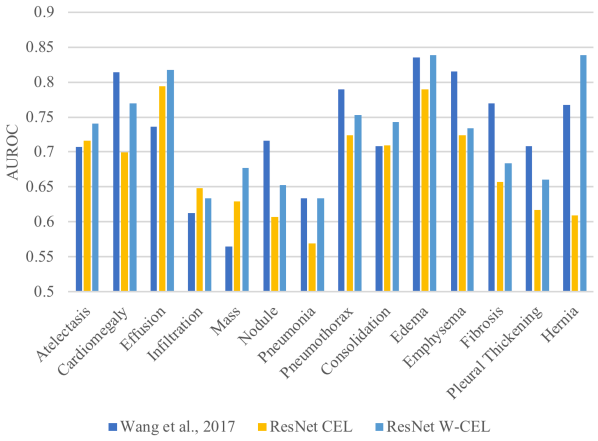
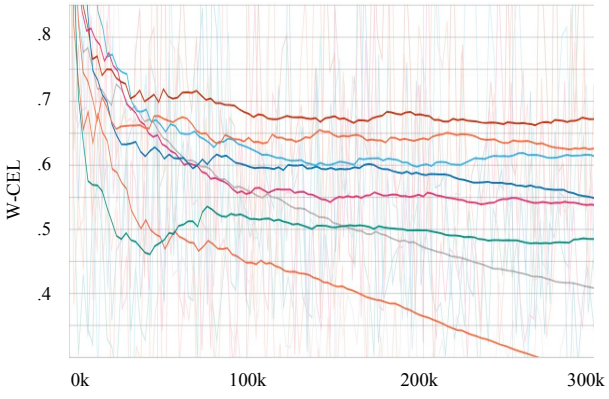


Figure 2. Comparison of ResNet to Wang et al., 2017. Using weighted cross entropy loss (W-CEL) increases the AUROC for less common diseases.



Color	Optimizer	Learning Rate Decay	Data Augmentation	Average AUROC
	Adam	Gradual	Flipping and Cropping	**.785**
	Momentum	Gradual	Flipping and Cropping	.753
	Adam	Step	Flipping and Cropping	.761
	Momentum	Step	Flipping and Cropping	**.790**
	Adam	Step	Flipping	**.788**
	Momentum	Step	Flipping	**.788**
	Adam	Gradual	Flipping	.633
	Momentum	Gradual	Flipping	.610

Figure 3. Results for DenseNet with different optimizers, learning rate decay schedules, and data augmentation strategies. The W-CEL is calculated with the training set and the AUROC is calculated with the test set.

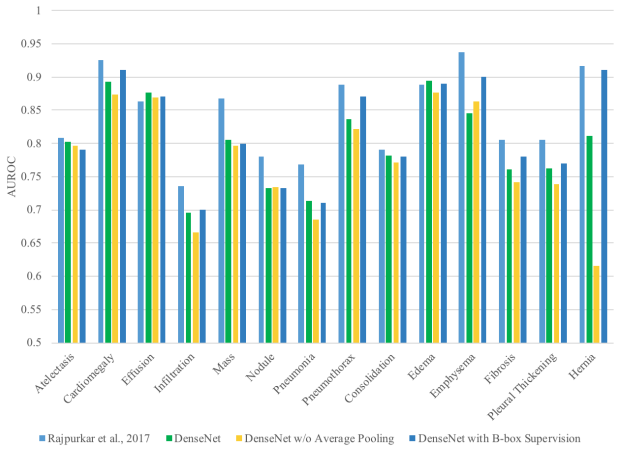


Figure 4. Comparison of DenseNet to Rajpurkar et al., 2017. Neither removing the average pooling layers nor supervising with bounding boxes effects the AUROC. The AUROC is calculated with the test set.

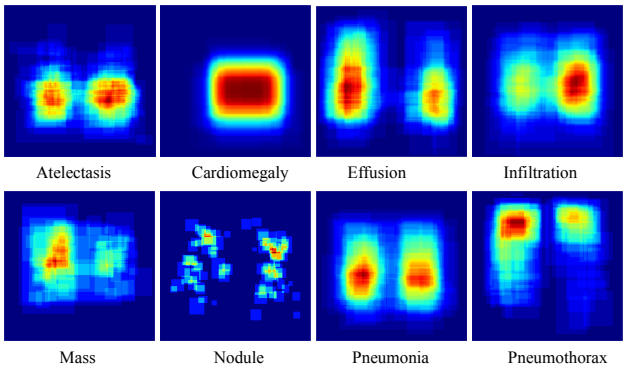


Figure 6. Spatial distribution of the bounding boxes across different diseases.

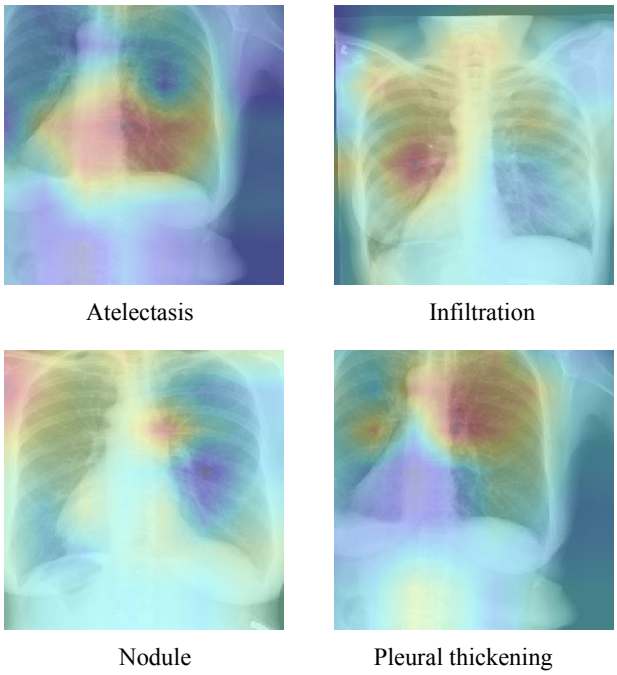


Figure 5. Example CAMs generated by DenseNet without bounding box supervision.

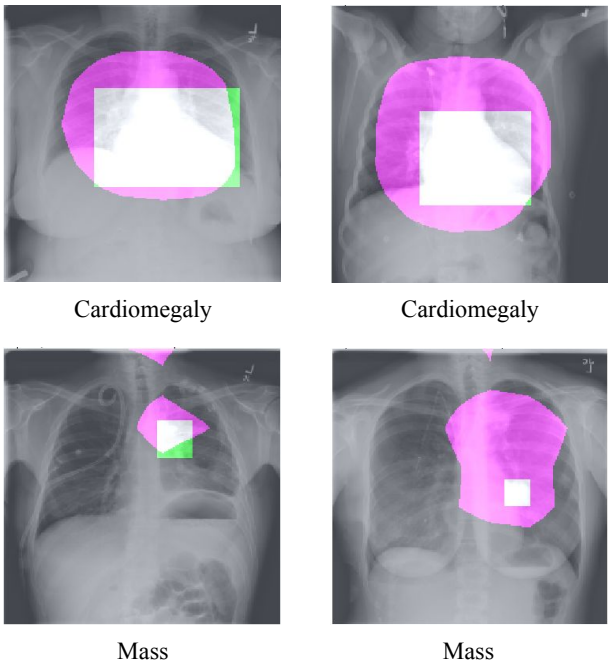


Figure 7. Localizations generated by bounding box supervised DenseNet compared to ground truth bounding boxes. Images are from the test set.

T(IoU)	Model	Atelectasis	Cardiomegaly	Effusion	Infiltration	Mass	Nodule	Pneumonia	Pneumothorax
.1	Baseline	**.69**	.94	**.66**	.71	**.40**	**.14**	.63	.38
.1	DenseNet	.20	.76	.47	.28	.22	0	.17	.37
.1	DenseNet w/ B-boxes	.29	**.98**	.45	.76	.17	0	.2	.53
.1	DenseNet w/o pooling	.41	1	.54	**.83**	.35	.08	**.75**	**.54**
.2	Baseline	**.47**	.68	**.45**	**.48**	**.26**	**.05**	.35	.23
.2	DenseNet	.02	.76	.22	.17	.11	0	.07	.16
.2	DenseNet w/ B-boxes	.07	**.97**	.25	.41	.17	0	.1	.21
.2	DenseNet w/o pooling	.05	1	.23	.44	.14	0	**.4**	**.38**
.3	Baseline	**.24**	.46	**.30**	.28	**.15**	**.04**	.17	**.13**
.3	DenseNet	.02	.74	.09	.14	0	0	.03	.05
.3	DenseNet w/ B-boxes	0	.94	.15	.31	.05	0	.07	.10
.3	DenseNet w/o pooling	0	**.94**	.06	**.31**	0	0	**.34**	.08
.4	Baseline	**.09**	.28	**.20**	.12	**.07**	.01	**.08**	**.07**
.4	DenseNet	0	.38	.03	.14	0	0	.03	0
.4	DenseNet w/ B-boxes	0	.74	.03	**.17**	0	0	.03	0
.4	DenseNet w/o pooling	0	**.76**	.04	.1	0	0	.07	0
.5	Baseline	**.05**	.18	**.11**	.07	**.01**	.01	**.03**	**.03**
.5	DenseNet	0	.14	.03	.1	0	0	.03	0
.5	DenseNet w/ B-boxes	0	.32	0	**0.13**	0	0	0.3	0
.5	DenseNet w/o pooling	0	**.40**	0	.1	0	0	0.3	0

Figure 8. Accuracy at IoU = [.1,.2,.3,.4,.5] for Wang et al., 2017, DenseNet, DenseNet with bounding box supervision, and DenseNet without average pooling

supervision from the bounding boxes like the lower resolution model. This yielded a slightly lower AUROC but significantly increased the IoU (figures, 4,8). The IoU values are comparable to the baseline established by Wang et al. However, Li et al. achieved significantly better localization. This is likely because I have not tuned the weight of the localization loss. Li et al scaled the localization loss by 5 since there are only a few bounding boxes (Li et al., 2017). I also have not done 5-fold cross validation, which is probably why the IoU for nodule is really low. There are only a few examples of nodules in the test set and I am getting them all wrong.

6. Conclusions

In conclusion, I achieve near state of the art identification of thoracic diseases by training a 121-layer DenseNet on the chestX-ray dataset (Rajpurkar et al., 2017). This task is formulated as a multilabel classification problem and I use a weighted loss function to insure the that the model learns to identify rare diseases. I achieve similar localization results to Wang et al without an post processing (Wang et al., 2017). I do this by training the model with limited supervision from the bounding boxes and removing the 2 x 2 average pooling layers to increase the resolution. My IoU values are generally low and do not compare to the state of the art.

7. Future Directions

This project is far from complete. My results so far provide a framework for the development of more complex models. Much of what I have done is similar to previous papers. In the future, I will focus on the increasing the ability of the model to localize diseases, since this is essential for computer aided diagnosis. I will do this by tuning the weight of the localization loss and possibly using a conditioned random field to iteratively generate ground truth bounding boxes for images without bounding boxes during training. Results similar to supervised localization have been reported using this technique (Li et al., 2018). I will also work on increasing the resolution of the heatmap without significantly increasing the training time. This could be accomplished using a U-Net which takes the low resolution heatmap and the scales it back up using deconvolution layers and skip connecting to higher resolution feature maps (Ronneberger et al., 2015). The ultimate goal is to have a network which can accurately localize small abnormalities, like nodules.

Acknowledgments

I would like to thank Ameet Soni for mentorship and guidance throughout this project me and Jeff Knerr his technical support.

References

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL http://arxiv.org/abs/1512.03385.

Huang, Gao, Liu, Zhuang, and Weinberger, Kilian Q. Densely connected convolutional networks. *CoRR*, abs/1608.06993, 2016. URL http://arxiv.org/abs/1608.06993.

Li, Guanbin, Xie, Yuan, and Lin, Liang. Weakly supervised salient object detection using image labels. *CoRR*, abs/1803.06503, 2018. URL http://arxiv.org/abs/1803.06503.

Li, Zhe, Wang, Chong, Han, Mei, Xue, Yuan, Wei, Wei, Li, Li-Jia, and Li, Fei-Fei. Thoracic disease identification and localization with limited supervision. *CoRR*, abs/1711.06373, 2017. URL http://arxiv.org/abs/1711.06373.

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

Ronneberger, Olaf, Fischer, Philipp, and Brox, Thomas. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015. URL http://arxiv.org/abs/1505.04597.

Wang, Xiaosong, Peng, Yifan, Lu, Le, Lu, Zhiyong, Bagheri, Mohammadhadi, and Summers, Ronald M. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *CoRR*, abs/1705.02315, 2017. URL http://arxiv.org/abs/1705.02315.

Yao, Li, Poblenz, Eric, Dagunts, Dmitry, Covington, Ben, Bernard, Devon, and Lyman, Kevin. Learning to diagnose from scratch by exploiting dependencies among labels. *CoRR*, abs/1710.10501, 2017. URL http://arxiv.org/abs/1710.10501.

Zhou, Bolei, Khosla, Aditya, Lapedriza, Àgata, Oliva, Aude, and Torralba, Antonio. Learning deep features for discriminative localization. *CoRR*, abs/1512.04150, 2015. URL http://arxiv.org/abs/1512.04150.