X-ray Thoracic Disease Classification and Localization Using Deep Neural Networks

Moein Sorkhei
sorkhei@kth.se

Abgeiba Yaroslava Isunza Navarro
ayin@kth.se

Alexander Nöu
anou@kth.se

Lingxi Xiong
lingxi@kth.se

Abstract

This project deals with the tasks of classification and localization of thoracic diseases using the newly released dataset of X-ray images called ChestX-ray14 [1] which contains 112,120 labelled chest X-ray images. We first try to replicate the baseline results of the paper, and then try to incorporate other methods containing the attention mechanism to see if it can improve the performance. Our reference paper for producing the results is [1].

1 Introduction

Deep learning has revolutionized image classification over the last decade. Its success has introduced computer vision techniques to a wide range of fields to improve, automate and support existing methods and create new ones. Particularly, in the medical field, image classification has been used for assisting the diagnosis of diseases through the analysis of medical images taken from different parts of the human body. Automating image analysis for medical purposes has become a very important technique to support clinical diagnosis. However, the performance of those automating models are not yet comparable to real doctors in order to offer accurate prediction, which leaves a lot of room for improvement. In this project, we focus on the diagnosis of diseases through chest X-ray images. Our main objective is to classify a set of chest X-ray images according to common thorax diseases using deep learning techniques, as well as localizing the disease areas. The research question throughout this project is how the attention mechanism can improve deep neural networks (DNN) for classification and localization of thorax diseases. We implement two types of attention mechanisms on top of our base model, namely Space attention and Attention gated models [2, 3] and investigate their effects on classification and localization of thorax diseases.

2 Related work

The original paper that collected the Chest X-ray database implement a baseline model for it, where four standard deep convolutional neural network, AlexNet, GoogLeNet, VGG-16, and ResNet-50, pretrianed on ImageNet, (please cite the corresponding papers for each of these models and also ImageNet (some of them are already in the ref.bib file) are used for predicting and localization tasks. We extend the work by using attention mechanism to guide the network to learn better, which also theoretically can improve the accuracy of localization.

Attention is a mechanism that guides the predictions by indicating where the important parts of the features are for making a certain prediction. Attention has been adapted to

various areas of machine learning and has helped improve the state of the art in fields such as computer vision [4, 5], natural language processing [6] and knowledge distillation [7]. Wang et al. [8] proposed Residual Attention Networks for more general image classification and attention residual learning for training very deep networks, where soft attention is incorporated within the residual architecture where it not only serves as a feature selector during forward inference but also act as a gradient update filter during back propagation. Lyu et al. [9] designed a global/local attention method that take the information from coarse to fine level feature to improve multi-label image classification.

Various approaches are tried on the Chest X-ray data as well. Huang et al. [10] designed a multi-attention convolutional neural network to learn a set of discriminate features for each category. Ma et al. [11] proposed a multi-attention convolutional neural network using the squeeze-and-excitation (SE) block as a feature attention module to re-calibrate cross-channel feature.

3 Problem description

The dataset contains 112,120 frontal chest X-ray images, each of which may belong to a single or multiple disease classes, and there are 14 different disease classes in total. Hence the problem is multi-label disease classification. The majority of the images in the dataset, however, do not contain any diseases which make the data highly imbalanced. Note that the labels corresponding to an image are represented by a vector whose length is equal to the total number of disease classes, and 1's and 0's in a vector indicate the presence or absence of the corresponding disease respectively. Hence, a vector of full 0's belongs to a normal case where no disease is observed in the image.

One way to tackle the problem of imbalanced dataset is to penalize the model for wrong predictions of the minority classes. This could be done by adding the appropriate coefficients to the usual cross-entropy loss (CEL) function. Hence the new loss equation would be the following [1], named as weighted CEL (WCEL):

$$L_{WCEL}(\mathbf{p}, \mathbf{y}) = \beta_P \sum_{y_c=1} -\ln(\mathbf{p_c}) + \beta_N \sum_{y_c=0} -\ln(1 - \mathbf{p_c}), \tag{1}$$

where $\mathbf{p}$ and $\mathbf{y}$ are the prediction and true label vectors respectively, β_P is $\frac{|P|+|N|}{|P|}$ and β_N is $\frac{|P|+|N|}{|N|}$. $|P|$ and $|N|$ are the number of 1's and 0's in a set of image labels respectively. Equation 1 is interesting as it takes into account how frequent different types of diseases are in a batch of label vectors. Since 0's in a label vector indicate the absence of the corresponding disease, β_P and β_N are adjusted in a way to give more importance to the disease class with fewer samples available.

3.1 Baseline model

The baseline model includes using a ResNet [12] pre-trained on the ImageNet [13] dataset and adding a transition layer, pooling layer, and prediction layer after that. The transition layer is a 1×1 convolutional layer applied to the feature maps (output of the ResNet) and can manipulate the channels of the feature maps. The output of this layer is passed through the pooling layer to obtain the most salient features, and finally a fully connected layer is applied to obtain the prediction vector (the prediction layer).

3.2 Space attention model

Inspired by [2] and [14] we approach the multi-label classification problem with a space attention mechanism. For a fair comparison with the Baseline model, we use the pre-trained ResNet on the ImageNet dataset and add an attention module on top of it. The attention module allows to add global context over the feature map from the ResNet without introducing a large computational cost. As seen in Figure 1, the feature map from the last layer of the ResNet is pooled over the entire image which results in a context vector. This vector is

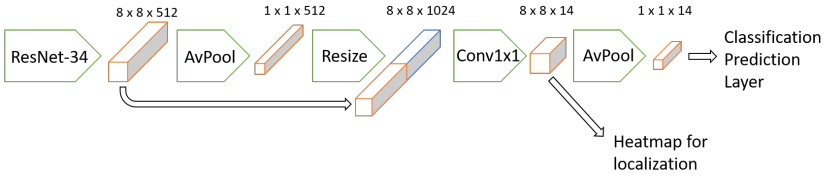

Figure 1: The overall flow-chart of the space attention network for classification and localization.

resized and appended to each of the original feature map. The resulting vector is introduced to a convolution to combine the global and local information into a feature vector of size $8 \times 8 \times n\_classes$. The resulting vector exploits channel dependencies which is then used directly for localization. For classification an extra average polling is applied which extracts the global information needed for classification.

3.3 Attention gated network

For the second attention mechanism, we adopted the attention gated model from [3]. As they claim that their design can specially fit medical image, where regions of interest are small, extremely localised and often highly non-homogeneous. Also the design can be easily added on an existing classification architecture. The structure of the pre-trained ResNet remains unchanged, two attention units are inserted after the second and third residual blocks, as shown in Figure 2. The attention gated modules (AG) provides the attention map of corresponding feature scales as well as a compatibility score between the current layer feature (fine-scale) and final layer feature (coarse-scale). The compatibility score between activation map of a chosen layer $\boldsymbol{f_i}$ and global feature vector $\boldsymbol{g}$ in channel s (becomes pixel-wise feature vector $\boldsymbol{f_i^s}$ of length C_s) is defined in Equation 2:

$$c_i^s = \boldsymbol{\Psi}\sigma_1(\boldsymbol{W_f f_i^s} + \boldsymbol{W_g g} + \boldsymbol{b_g}) + \boldsymbol{b}_\psi \tag{2}$$

where $\boldsymbol{\Psi}$ is a learnable parameter, $\boldsymbol{W_g}$ is a learnable weight to match the dimension of $\boldsymbol{g}$ to $\boldsymbol{W_f f_i^s}$. The attention mechanism used here is more general, where $\boldsymbol{W_f} \in \mathbb{R}^{C_{int} \times C_s}$, non-linearity σ_1 (ReLU) cite ? and two bias terms $\boldsymbol{b}_\psi \in \mathbb{R}$, $\boldsymbol{b_g} \in \mathbb{R}^{C_{int}}$ are introduced as well. The usage of W_f, which introduces an intermediate channel to the fine-scale layer, has two benefits. First is to "weaken" the fine-scale (early) layer to generate a single signal that is compatible to $\boldsymbol{g}$ while helping with learning discriminant features. Second, the $\boldsymbol{W_f}$, $\boldsymbol{W_g}$, and σ_1 enable the network to more expressively learn non-linear relationships between vectors, since the medical images are inherently noisy and the region of interest is non-homogeneous.

Since there are multiple attention maps at different scales, aggregation is needed in order to provide the final prediction. In this project we used the most standard way which is to concatenate the scores from all three levels and feed them into a linear classifier.

4 Implementation details

In this project, we used the PyTorch framework [15] for implementing the neural networks, and to make the experiments more feasible, we down-sampled each image from 1024×1024 to the size of 256×256. Each image is normalized according to the mean and standard deviation from the training set of ImageNet [13]. We used the same data split provided by [1] with three subsets: 10% for validation, 20% for testing, and the rest for training. In addition, 984 test samples include bounding boxes manually identified by radiologists (for the original 8 classes of the dataset), for testing the localization ability of our models. Training data only contains disease labels, meaning that the bounding boxes are only used at test time, and hence localization task is weakly supervised. Our code could be found at: `https://github.com/AbIsuNav/Data_Science_Project`.

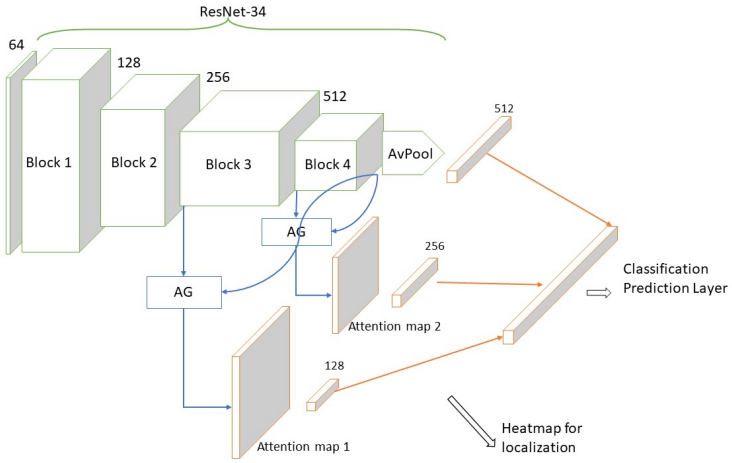

Figure 2: The overall flow-chart of the attention gated network for classification and localization (aggregate mode is concatenate).

4.1 Baseline model

For the baseline model, we decided to use the ResNet-34 [12] pre-trained on the ImageNet dataset [13] as the baseline network. We removed the final pooling and fully connected layers and added the transition, pooling, and prediction layers to the output of the network. The authors of the reference paper [1] decided to freeze the ResNet layers (they used ResNet-50) and only train the other layers. In our experiments, we both freezed and unfreezed the layers and observed that when the ResNet layers are not freezed, the classification results improve significantly, but at the cost of longer training time per epoch, as expected. Using such dimensions for the images, the output of the ResNet model would be of size $8\times8\times512$. Next, the output of the ResNet is passed through the transition layer, a 1×1 convolutional layer. It is then passed through the max-pooling layer to get the output shape of $1\times1\times512$ and finally through the fully connected layer to produce probabilities (using the Sigmoid function) for the image belonging to each of the 14 classes. In order to find a good learning rate, we ran grid search over 10 different values ranging from 0.001 to 1e-6, and based on the validation loss evaluation, we decided to choose 1e-4 as the appropriate learning rate. Then we trained the network with the Adam optimizer [16] for 30 epochs with early stopping, terminating the training if no improvement in validation loss is observed after 3 consecutive epochs. We used mini-batches of size 128 and trained the model using an Nvidia Tesla K80 GPU.

For the localization task, we followed the procedure of [1] and extracted the final feature maps of the network before the pooling layer. The feature maps are then multiplied with the fully connected classifier layer in order to create heat-maps for each class. The heat-maps indicate which areas are important for classifying the corresponding disease, and can therefore be used for the localization task. A simple threshold method, with threshold 180, is used to generate bounding boxes of the believed disease area Can you elaborate what this threshold means?.

4.2 Space attention model

For the space attention we followed a similar approach as in the baseline model by using the pre-trained ResNet-34 [12] on the ImageNet dataset to initialize the network and removing the final pooling and fully connected layers. However, as shown in Figure 1 we applied an average pooling to the feature maps from the ResNet-34. Then, we resized the vector obtained from the polling layer and concatenated it with the original feature maps from the ResNet. We extracted the information from both images using a single convolution over. Its output is a vector size $8 \times 8 \times 14$ which we used for the localization task. For the

classification layer we extracted the global information using an average polling layer and finally we obtained the probabilities for the multi-classification through a Sigmoid function. The Adam algorithm [16] was adopted to optimize the network parameters by minimizing WCEL with weight decay of 0.0001. The initial learning rate was set to 1e-4 and divided by 10 every 20 epochs. An early stopping condition was adopted to avoid over-fitting the model. As in the baseline model, the training of the model was stopped after an increment of the validation loss in 3 consecutive epochs. We used the batch size of 128 and an Nvidia Tesla P100 GPU for training this model.

4.3 Attention gated model

The $\boldsymbol{W_f}$, and $\{\boldsymbol{\Psi}, \boldsymbol{b_\psi}\}$ in Equation 2 are implemented as 1×1 convolution layer with stride=(1, 1). And $\{\boldsymbol{W_g}, \boldsymbol{b_g}\}$ is a fully connected layer. The upsample method is bi-linear. We used it to match the coarse gird to spatial resolution in all situations. For training, we used the Adam optimizer with a learning rate of 0.0001. For classification, the compatibility scores and the output of final residual block after pooling are concatenated and a fully connected layer with Sigmoid activation is used as a linear classifier. For localization testing, the attention map of different levels(from block 2 and block 3) are extracted together with the final feature before pooling. We normalize and take the mean of them to put into the heat-map and bounding box generating module we described before. We used the batch size of 32 and an Nvidia GeForce GTX1060 GPU for training this model

5 Experimental results

5.1 Disease classification

Since the problem is multi-label classification, AUC values of the ROC curves are used to measure the capability of the trained model to identify the presence of each disease in the test data. Pleas note that in all the tables (including both classification and localization), the numbers reported from the reference paper are actually the results of using images of size 1024×1024, while our results are based on images of size 256×256.

Baseline model. Here we report the test results of our best baseline model (with the unfreezed ResNet). The ROC curves for different disease classes can be observed in Fig. 3 and the corresponding AUC values are reported in Table 1 in which we also bring the baseline results of the original paper [1] for comparison. As we can see, the AUC values of our model closely follow or outperform the numbers reported by the authors.

Space attention model. The results from the Space attention model are shown in table 1 and the ROC curve in Figure 4. The results are derived from the model with the lower validation error. The model outperforms in the classification the Baseline ResNet-50 model in all classes and the Baseline ResNet-34 model in 10 out of 14 diseases. The heat-maps were able to locate the disease regions without the need of the model being trained for localization.

Attention gated network. The results of the attention gated network can be found 1, and the ROC curve is presented in 5.

5.2 Disease localization

The localization capabilities of the models are evaluated on the aforementioned test samples by comparing the generated bounding boxes with the ground truth boxes, using the metrics intersection over union (IoU) and intersection over bounding box (IoBB):

$$\text{IoU} = \frac{\text{Area of Intersection}}{\text{Area of Union}}, \tag{3}$$

$$\text{IoBB} = \frac{\text{Area of Intersection}}{\text{Area of Bounding Box}}. \tag{4}$$

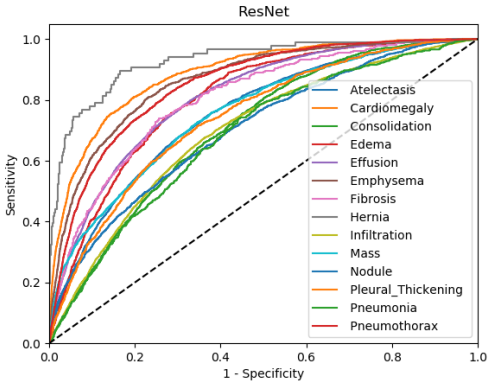

Figure 3: The ROC curve for different classes produced by our baseline model.

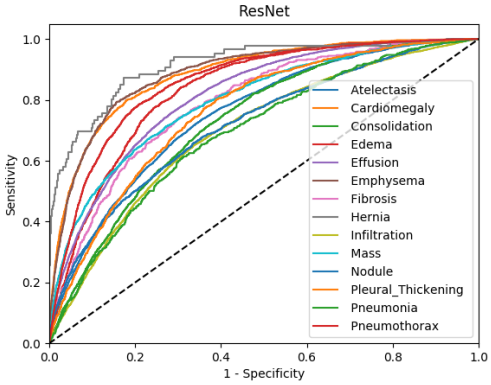

Figure 4: The ROC curve for different classes produced by our space attention model.

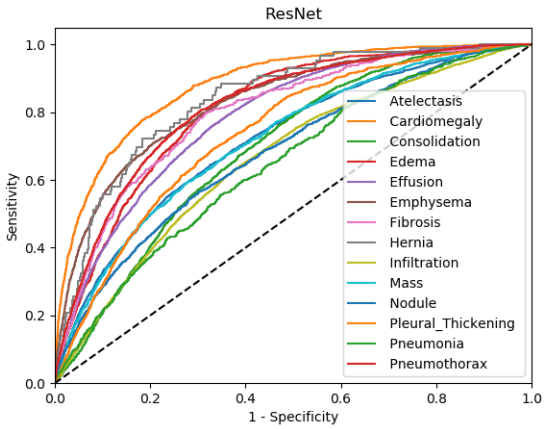

Figure 5: The ROC curve for different classes produced by our attention gated model.

Table 1: AUC results for different models. BL stands for baseline, S.Att. for space attention. The values of BL ResNet-50 are directly reported from [1].

Disease class	BL ResNet-50	BL ResNet-34 (ours)	S.Att. (ours)	AG (ours)
Atelectasis	0.7003	**0.7522**	0.7519	0.7180
Cardiomegaly	0.8100	**0.8856**	0.8842	0.8800
Effusion	0.7585	0.8091	**0.8122**	0.7860
Infiltration	0.6614	0.6978	**0.6979**	0.6620
Mass	0.6933	0.7563	**0.7956**	0.7178
Nodule	0.6687	0.7038	**0.7172**	0.6790
Pneumonia	0.6580	0.6830	**0.6923**	0.6460
Pneumothorax	0.7993	0.8501	**0.8570**	0.8147
Consolidation	0.7032	0.7056	**0.7230**	0.6932
Edema	0.8052	0.8032	**0.8314**	0.8072
Emphysema	0.8330	0.8632	**0.8860**	0.8278
Fibrosis	0.7859	**0.8032**	0.7881	0.7973
Pleural_Thickening	0.6835	0.74280	**0.7619**	0.7377
Hernia	0.8717	**0.9266**	0.9138	0.8376

As is mentioned in [1], due to the relatively low spatial dimensions of the heat maps ($\{8, 16, 32\}$) compared to the images (in our case down sampled to 256×256), the resulting bounding boxes are usually much larger when resized to match the images. We therefore compute the accuracy of successfully localized diseases for several thresholds, i.e. T(IoU) $= \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7\}$ and T(IoBB) $= \{0.1, 0.25, 0.5, 0.75, 0.9\}$. A bounding box is considered successful if IoU $>$ T(IoU) or IoB $>$ T(IoB).

Table 2 and 3 show results for one threshold of each metric. Both our models outperform the results of [1], and the baseline model has slightly higher accuracy than the space attention model. Note that this is just for one specific threshold, and smaller thresholds may favor larger, overestimating bounding boxes. Figure 6, 7 and 8 show qualitative examples of bounding boxes and heat-maps. More results for different thresholds will be found in the Appendix.

Table 2: Localization accuracy for different models, using threshold T(IoU) $= 0.3$. BL stands for baseline, S.Att. for space attention and AG for attention gated network. The values of BL ResNet-50 are directly reported from [1].

Disease class	BL ResNet-50	BL ResNet-34 (ours)	S.Att. (ours)	AG (ours)
Atelectasis	0.2444	0.5230	**0.5402**	0.3851
Cardiomegaly	0.4589	**0.9926**	0.9778	0.8889
Effusion	0.3006	0.2266	**0.3594**	0.2344
Infiltration	0.2764	**0.5806**	0.5081	0.5645
Mass	0.1529	**0.4149**	0.3511	0.2021
Nodule	0.0379	**0.0395**	0.0132	0.0263
Pneumonia	0.1666	0.3830	0.2979	**0.5106**
Pneumothorax	0.1326	**0.1429**	0.1339	0.0714

Table 3: Localization accuracy for different models, using threshold T(IoBB) $= 0.25$. BL stands for baseline, S.Att. for space attention and AG for attention gated network. The values of BL ResNet-50 are directly reported from [1].

Disease class	BL ResNet-50	BL ResNet-34 (ours)	S.Att. (ours)	AG (ours)
Atelectasis	0.5500	**0.8563**	0.8103	0.8276
Cardiomegaly	0.9794	0.9926	**1**	0.9630
Effusion	0.5424	**0.7813**	0.7344	0.7031
Infiltration	0.5772	**0.9194**	0.7742	0.8468
Mass	0.2823	**0.7872**	0.7766	0.7340
Nodule	0.0506	0.6316	**0.6842**	0.6184
Pneumonia	0.5583	**0.8652**	0.7943	0.8582
Pneumothorax	0.3469	0.3661	**0.4375**	0.3214

6 Discussion

In this work, we tried different methods for the task of medical images classification, and observed that in the baseline model, unfreezing the layers of the pre-trained ResNet would in fact improve the classification performance significantly. The attention mechanisms showed higher performance on classification for most of the diseases. We attribute the higher performance of the space attention model over the baseline models to the global and local information that is extracted from the feature maps. All the models implemented were able to intuitively localize the disease regions without the need of annotated regions or specialized train. Our results have indicated that the three models outperform on classification and localization the model with ResNet-50 implemented by [1].

It is also noteworthy, as discussed with the other group, that it would have been much more informative if we had run the experiments multiple times and provided error bars, since the randomness may play a role in test performance. Furthermore, although the test set was

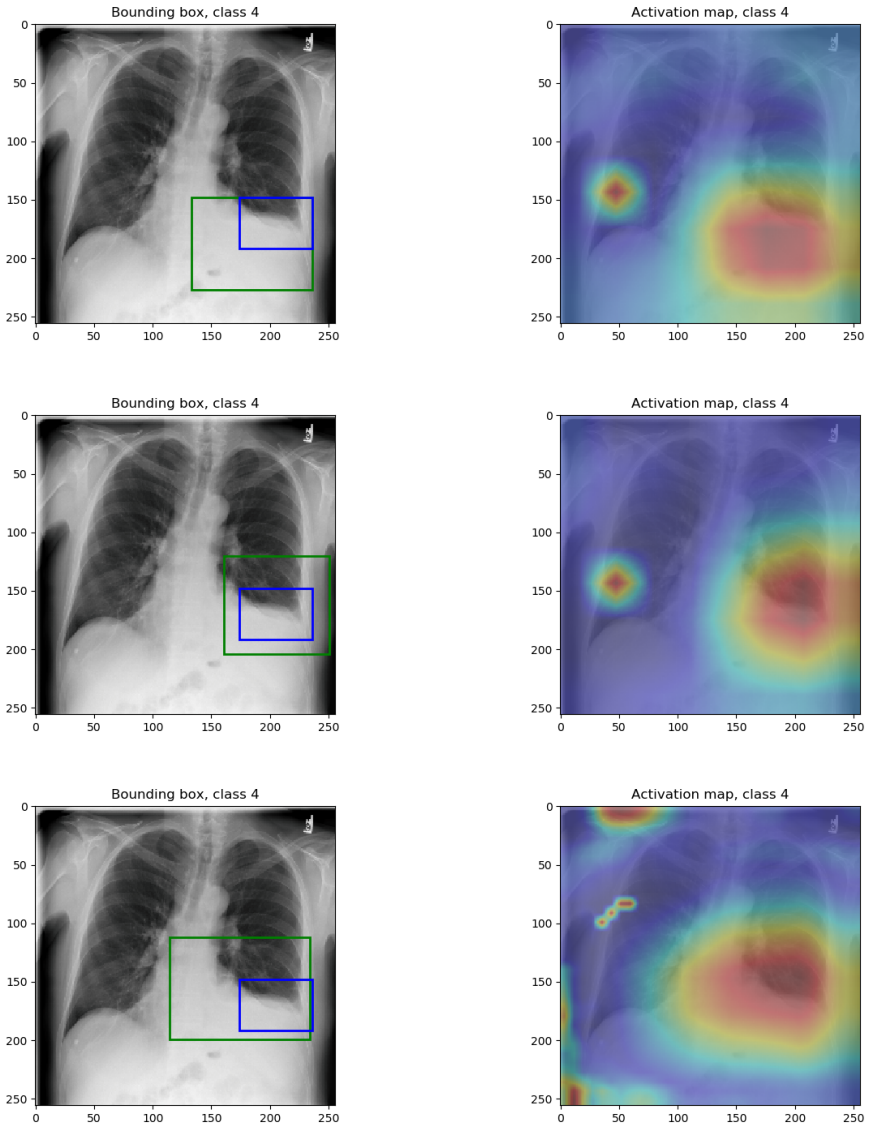

Figure 6: Visual examples of bounding boxes and heat-maps for the baseline, space attention and attention gated models (in this order by row) on a test sample of class Effusion. The blue box represents ground truth and the green box is the generated bounding box for the corresponding class.

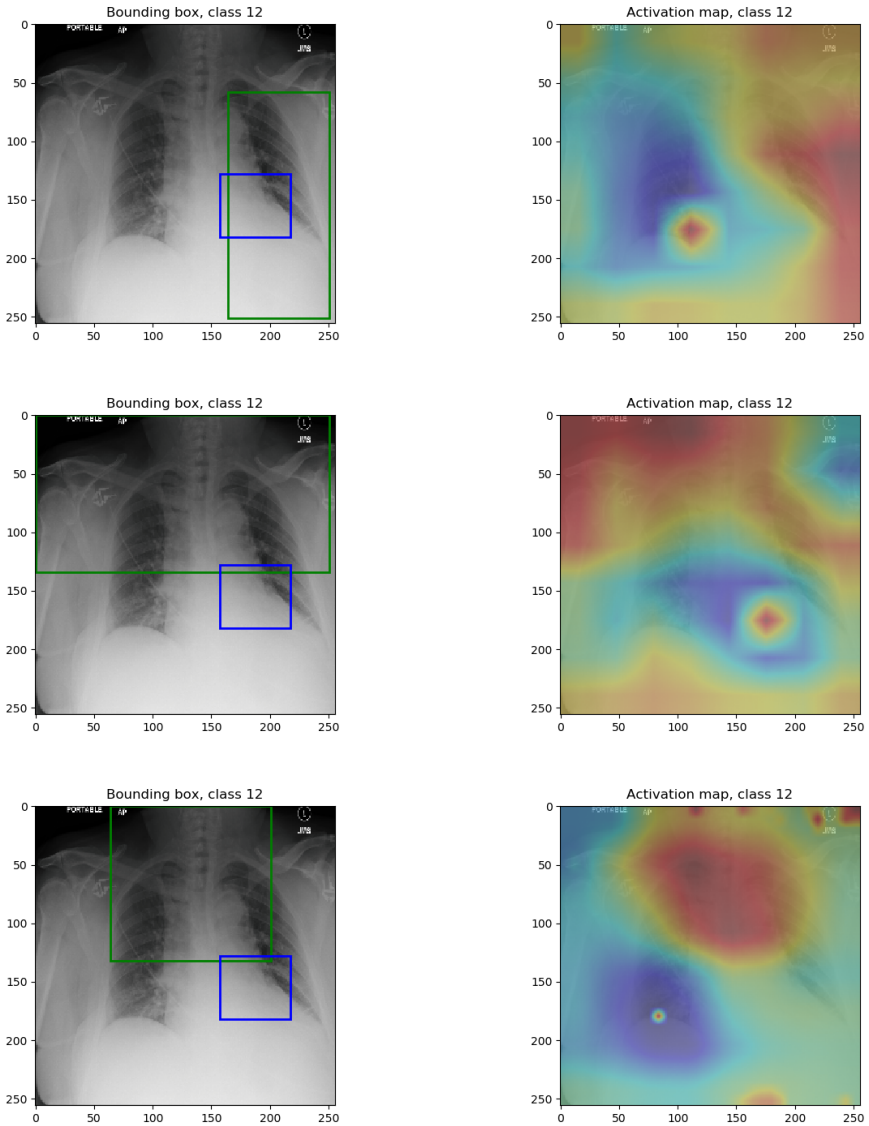

Figure 7: Visual examples of bounding boxes and heat-maps for baseline, space attention and attention gated models (in this order by row) on a test sample of class Pneumonia. The blue box represents ground truth and the green box is the generated bounding box for the corresponding class.

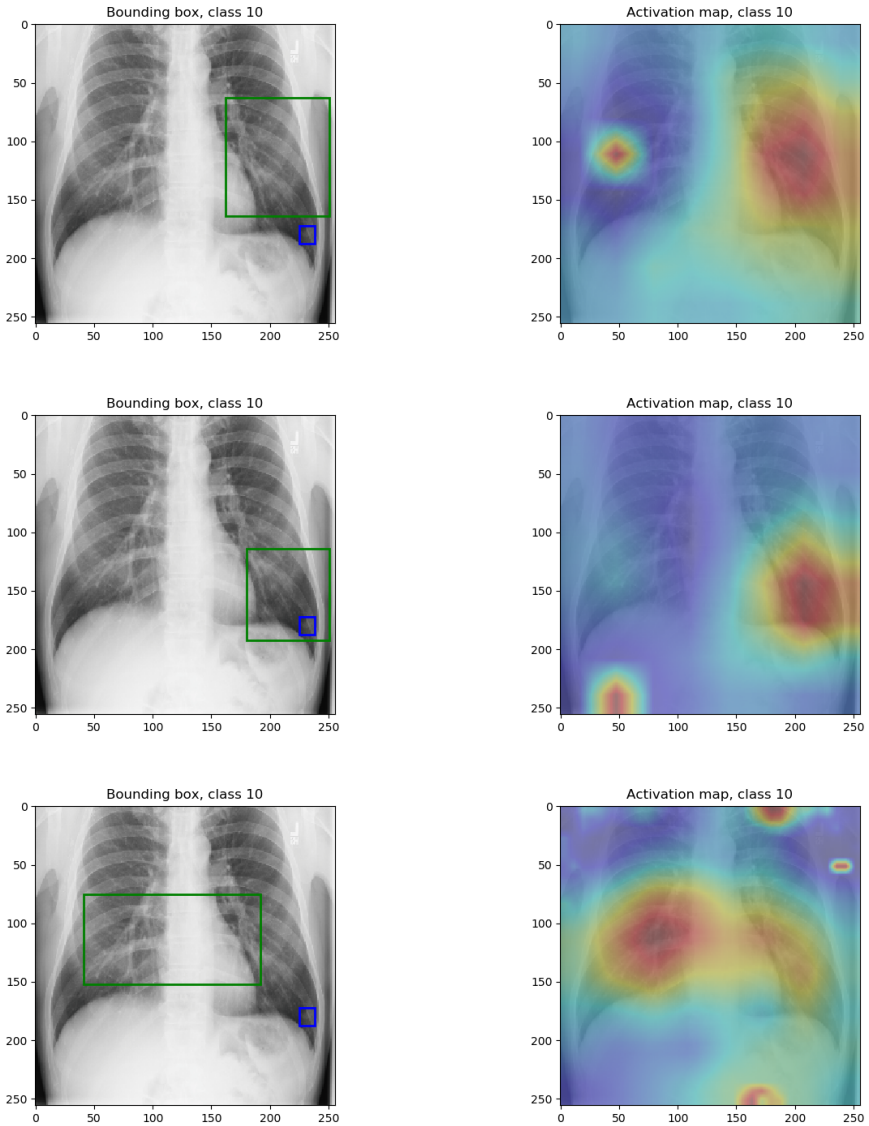

Figure 8: Visual examples of bounding boxes and heat-maps for baseline, space attention and attention gated models (in this order by row) on a test sample of class Nodule. The blue box represents ground truth and the green box is the generated bounding box for the corresponding class.

already determined by the dataset website, we randomly split the data (excluding the test images) in each run to obtain the training and validation sets. It would have been better if we had used the same random seed every time we split the training and validation sets to make sure all the models are trained and validated using exactly the same set of images.

Explain why the attention models were not significantly better for localization (or classification) which might essentially have to do with learning rate and the way they are trained.

Exaplin why attention-gated did not actually outperform our baseline model.

Discuss that, like in Figure 7, the localization could not be very acurate and may contain major parts of the image, part of which is due to the inherent difficulty in the localization task

Discussion with the other group (group 6). We discussed several items with the other group. The have been trying to replicate the results of [17]. They have focused on the classification of only one disease: Pneumonia. What they have done to further improve the classification results is to use the focal loss [18] which gives even more importance to hard examples in the data. The focal loss has been defined mainly for dense object localization, and seems a great choice for the problem at hand. At the time of writing this, they have been issues regarding the data: since their baseline method works based on an ensemble of ResNet models (plus some additional layers), it is designed to accept sets of images of different sizes: 128×128, 256×256, and 384×384. This makes their model quite heavy and data-intensive, and they had to reduce the batch size to 12 so the mini-batch could fit the GPU memory. It is good to add something if you remember how their data augmentation (channel shift) worked

References

[1] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

[2] Yanbo Ma, Qiuhao Zhou, Xuesong Chen, Haihua Lu, and Yong Zhao. Multi-attention network for thoracic disease classification and localization. pages 1378–1382, 05 2019.

[3] Jo Schlemper, Ozan Oktay, Liang Chen, Jacqueline Matthew, Caroline Knight, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention-gated networks for improving ultrasound scan plane detection. *arXiv preprint arXiv:1804.05338*, 2018.

[4] Saumya Jetley, Nicholas A Lord, Namhoon Lee, and Philip HS Torr. Learn to pay attention. *arXiv preprint arXiv:1804.02391*, 2018.

[5] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[7] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

[8] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2017.

[9] Fan Lyu, Fuyuan Hu, Victor S Sheng, Zhengtian Wu, Qiming Fu, and Baochuan Fu. Coarse to fine: Multi-label image classification with global/local attention. In *2018 IEEE International Smart Cities Conference (ISC2)*, pages 1–7. IEEE, 2018.

[10] Zhicheng Huang and Dongmei Fu. Diagnose chest pathology in x-ray images by learning multi-attention convolutional neural network. In *2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, pages 294–299. IEEE, 2019.

[11] Yanbo Ma, Qiuhao Zhou, Xuesong Chen, Haihua Lu, and Yong Zhao. Multi-attention network for thoracic disease classification and localization. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1378–1382. IEEE, 2019.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[13] 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.

[14] Wei Liu, Andrew Rabinovich, and Alexander C. Berg. Parsenet: Looking wider to see better, 2015.

[15] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[16] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014.

[17] Tae Joon Jun, Dohyeun Kim, and Daeyoung Kim. Automated diagnosis of pneumothorax using an ensemble of convolutional neural networks with multi-sized chest radiography images. *arXiv preprint arXiv:1804.06821*, 2018.

[18] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

Appendix

Here we list more quantitative localization results with differen thresholds for IoU and IoBB. It is great to have more localization visualizaitons as well, if you have the time!