Classification and Localization of Disease with Bounding Boxes from Chest X-Ray Images

Hugo Kitano

https://youtu.be/k7SxhheHW7A

Project Summary

For the final project, I trained a model that classifies and localizes disease from the NIH Chest X-Ray Database. The model was trained on classification on training and validation datasets, optimizing mean AUC over all eight classes. Then, the class activation maps of the images that have bounding-box data from the dataset were taken using Grad-CAM++. Finally, bounding boxes were retrieved from the activation maps, and evaluated using metrics such as IoU. I found that the questionable labeling of the dataset capped the accuracy of the model's classification, but still, both the classification and localization turned out to be useful.

Introduction

- Chest x-rays are the most common type of radiology exam in the world, but diagnosing one of the many possible chest afflictions to the many organs and systems in the chest is a difficult task.
- Most of the models that deal with the NIH Chest X-Ray Dataset deal with classification of chest disease from X-Ray images.
- However, this project is focused on both **classification and localization**. Approximately 1% of images in the dataset have their disease localization described by bounding boxes. Here is the general approach:
  - Train a multi-class classification model on the training and validation set, optimizing for the mean AUC over all classes.
  - Obtain the images' class activation maps (CAMs) using Grad-CAM++.
  - Procure bounding boxes from the CAMs.
  - Evaluate the results with intersection-over-union and other metrics.

Approach

Our approach has four key steps:

- **Pre-processing**: saving intermediate arrays with the image data makes training much faster. We resize our images from 1024x1024 to 256x256, and split our train/val set 90-10 (train/val and test sets are given in the data).
- **Training:** I used a DenseNet121 model, pre-trained on ImageNet, to do multi-class classification on the images. I used data augmentation (a random crop and a random horizontal flip) to increase the training data size, and used weighted loss to counteract class imbalance. For validation, I calculated the mean AUC over all classes to choose the best model.
- **Class Activation Maps**: I used Grad-CAM++, an extension of Grad-CAM that is better for object localization, to extract class activation maps from images. Each CAM is associated with a particular class. For every image with a bounding box, we collect all CAMs for where the class probability is above a certain threshold. This means we often collect multiple CAMs per image.
- **Bounding Boxes**: For each image with a bounding box, we collect all the class activation maps, and keep all activations greater than some value *t*. Then, we choose the largest connected rectangle remaining as the bounding box. I chose to favor larger bounding boxes, since the goal of this project is to provide a diagnostic for medical professionals..

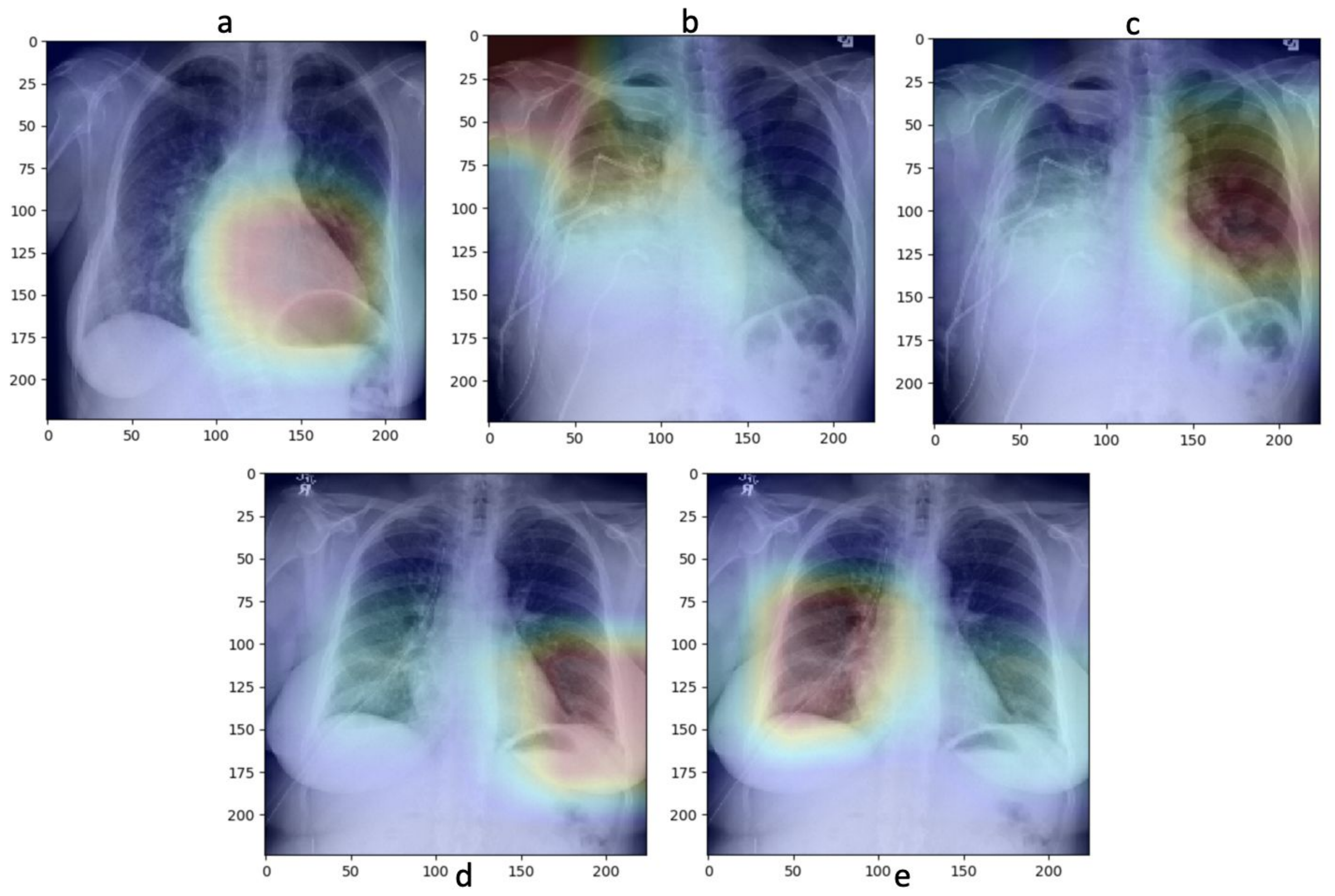
I based the first two steps my code on T.H. Tang's repository on Github[1] which implements the CheX-Net model. I used WonKwang Lee's repository on Github[2] for his implementation of Grad-CAM++.

Data

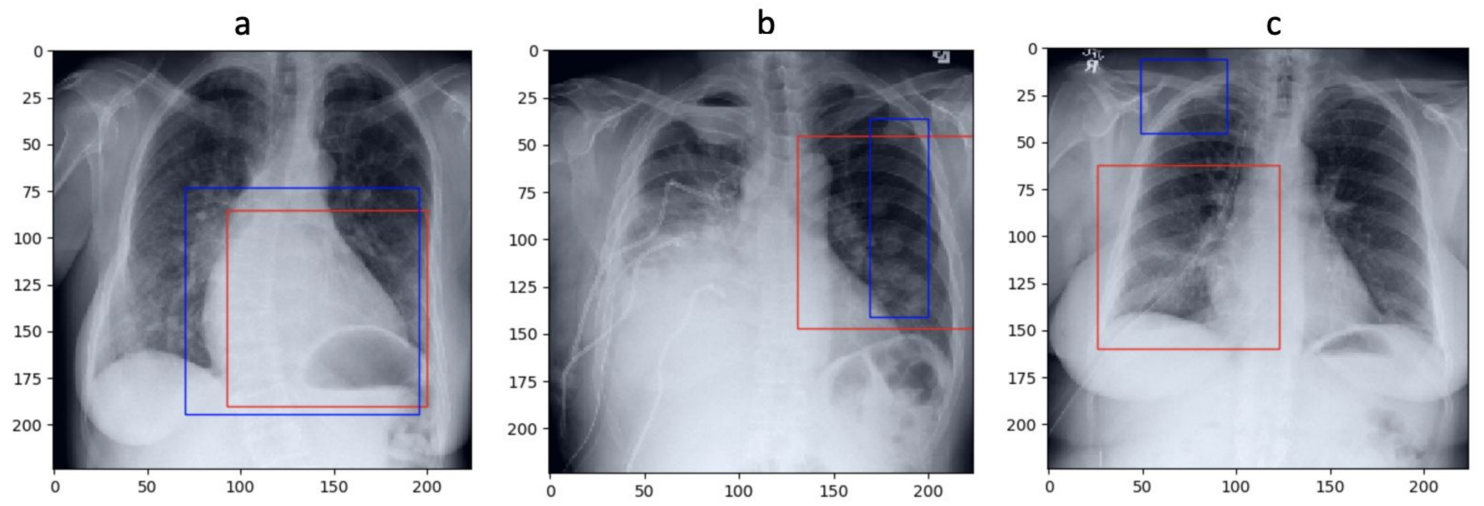
- The NIH Chest X-ray Dataset is comprised of 112,120 X-ray images with disease labels from 30,805 unique patients. It contains 14 disease classes, plus an additional "no findings" class. Images can be afflicted with multiple diseases.
- 983 images are described with bounding boxes (all part of the test set) by professional radiologists. 8 of the disease classes are represented here: atelectasis, cardiomegaly, effusion, infiltrate, mass, nodule, pneumonia, and pneumothorax.
- The bounding boxes are (x,y,w,h) tuples, where x and y locate the upper corner of a rectangle, and w and h describe the size of the rectangle.
- The labels are created not by humans, but by using NLP techniques with their corresponding radiology reports. They say their accuracy is 90%, but even that is in question.
- There's a lot of class imbalance, which makes training and evaluating sometimes difficult. We use a weighted loss and AUC metrics instead of accuracy to mitigate it.

Experiments

- Here are five class activation maps corresponding to three images
  - ***a*** is an image of cardiomegaly, with a cardiomegaly CAM.
  - ***b*** and ***c*** are images of mass, where ***b*** is a nodule CAM and ***c*** is a mass CAM.
  - ***d*** and ***e*** are images of pneurothorax, where ***d*** is an atelectasis CAM and ***e*** is a pneumonia CAM.



- Ground truth (blue) and prediction (red) bounding boxes overlaid over original image. ***a*** is an instance of cardiomegaly, ***b*** is of an instance of mass, and ***c*** is of an instance of pneumothorax.



Evaluation

- For classification training, the best DenseNet model had an **average AUROC of 0.779 on validation**, and **0.758 on test**. When limiting the test set to the images with bounding-boxes, the average AUROC is 0.784, which is quite high. This model used weighted loss.
- AUC values per class on the test set and the bounding-box test set. A wide variance!

	Atelectasis	Cardiomegaly	Effusion	Infiltrate	Mass	Nodule	Pneumonia	Pneumothorax
Test Set	0.747	0.867	0.813	0.602	0.804	0.738	0.677	0.810
Test Bounding Box Set	0.789	0.884	0.848	0.629	0.822	0.827	0.664	0.811

- Three metrics of success:
  - **intersection-over-union (IoU)** measures the intersection of two bounding boxes and divides that by the union of the two bounding boxes
  - **containment** measures whether one of the bounding boxes completely contains the other.
  - **non-overlap** describes when neither boxes overlap with each other (and have an IoU of zero). These are clear failures.
- The final Grad-CAM++ model has an **average IoU of 0.201**, with a **19.3% non-overlap rate** and a **35.4% containment rate**. It clearly outperforms a Grad-CAM implementation, which has an average IoU of 0.186, a 21.4% non-overlap rate and a 32.8% containment rate.
- Number of images, average IoU, non-overlap, and containment per class:

	Atelectasis	Cardiomegaly	Effusion	Infiltrate	Mass	Nodule	Pneumonia	Pneumothorax
Test Images	180	146	153	123	85	79	120	98
IoU	0.102	0.534	0.174	0.227	0.124	0.013	0.232	0.083
Non-overlap	0.194	0.007	0.222	0.114	0.165	0.443	0.100	0.459
Containment	0.538	0.055	0.261	0.398	0.612	0.443	0.375	0.224

Conclusions and Next Steps

- I was able to train a model for both classification and localization, dealing with an unbalanced dataset, clear data inaccuracies, and no bounding boxes to train with. Yet, the model classified decently with strong AUC, and was able to localize many of the classes, such as cardiomegaly and infiltrate, well. For most classes, my choice of larger bounding box seems to work well to indicate the region of disease.
- Some possible adjustments:
  - different train-validation split: because of the data imbalance, different splits can affect the model greatly
  - trained with a smaller training rate (or with learning rate decay) to try to pinpoint the model with the best validation
- More accurate training dataset is a must: if we could be more sure about our classification labels, we could use fewer CAMs and then use medical disease information to design our boxes.

References

- [1] Tang, T.H. "Weakly Supervised Learning for Findings Detection in Medical Images." GitHub, 7 Aug. 2019, github.com/thtang/CheXNet-with-localization.
- [2] Lee, WonKwang. A Simple Pytorch implementation of Grad-CAM, and Grad-CAM++. GitHub, 3 Aug. 2018, https://github.com/1Konny/gradcam_plus_plus-pytorch.