Chest Abnormalities Detection and Localization - Progress Report

Kanstantsin Pachkouski
Electrical and Computer Engineering
Western University
London, Canada
kpachkou@uwo.ca

Kyle Rioux
Electrical and Computer Engineering
Western university
London, Canada
krioux5@uwo.ca

Christopher Tam
Electrical and Computer Engineering
Western University
London, Canada
ctam86@uwo.ca

***Abstract*—Current automated approaches for examining chest X-rays lack accuracy and list only entire-image findings without descriptive localization. An automated method to accurately identify chest X-ray abnormalities with localization in the image would therefore be an invaluable tool in providing improved healthcare globally. Thus, we aim to develop a computer-aided tool that is able to perform accurate detection of 14 common chest X-ray abnormalities with bounding boxes. In addition, each detected abnormality will have a confidence percentage associated with it.**

I. Introduction

Diagnostic medical imaging is an invaluable tool in the treatment of patients around the world. The development of imaging modalities such as CT, MRI, PET, and X-ray have allowed healthcare professionals to gain critical information which would be otherwise unavailable or require an invasive procedure. Although some images are relatively easy to interpret and can be done by a general practitioner, in many cases, it takes a radiologist to draw diagnosis from the complex images. It is well known that there is a global radiologist shortage, especially in developing countries [1]. The development of computer programs to analyze complex medical images could greatly reduce the burden on healthcare systems all around the world and lead to improved patient outcomes.

Among the most complex and common medical images is the chest X-ray. As X-rays are far cheaper than alternative imaging modalities such as CT, MRI, or PET, they are much more commonly used in developing countries. As the chest X-ray includes a large portion of the body, there are many conditions which are identified utilizing the chest X-ray. Current automated chest X-ray interpretation techniques are able to identify what they believe is present in an image, but give no localization as to where they believe the finding is within the image; this leads to a low level of interoperability and could add to the confusion of a doctor reviewing findings of the automated system.

We aim to develop a computer-aided diagnostics system which is able to perform detection on 14 common findings within a chest x-ray by localizing them in the image with a region of interest (ROI) bounding box, as well as a confidence percentage. For example, the system may provide the output similar to the following: 'The system is 70% sure that there is a nodule in this location'. This additional information will provide increased interpretability to computer-aided diagnostics systems and will allow them to fit more readily into current clinical workflows.

II. Related Work

Previous work in object detection tasks with localization have used R-CNN, or Regions with CNN Features networks. R-CNN is an object detection model that applies high-capacity CNNs to bottom-up region proposals in order to localize and segment objects [2]. R-CNN uses selective search to identify a number of bounding-box object region candidates, and extracts features from each region independently for classification. This novel approach of combining region proposals with CNNs heavily outperformed previous best performing models which relied on combining multiple low-level image features with high-level context. R-CNN gave a 30% relative improvement over the best previous results on the Pascal Visual Object Classes Challenge 2012 (PASCAL VOC 20212) competition.

In 2015, Girshick proposed Fast R-CNN [3], an improvement over R-CNN by aggregating regions of interest into a single forward pass over the image. In this context, regions of interest from the same image share computation in the forward and backward passes. Specifically, Fast R-CNN improves R-CNN in the areas of detection quality, training speed, and memory requirements. Fast R-CNN was again evaluated on the VOC 2012 dataset and improved over it's predecessor by 6%, achieving a mean average precision (mAP) of 65.7%. Fast R-CNN was also two orders of magnitude faster than other methods based on the comparably slower R-CNN pipeline.

Later, Faster R-CNN [4] was proposed as an improvement over Fast R-CNN. Faster R-CNN introduces a Region Proposal Network (RPN) that shares full-image convolutional features with the detection network. RPN is a fully convolutional network that simultaneously predicts object bounds with their associated probabilities at each position, thus enabling nearly cost-free region proposals. RPN is used in conjunction with Fast R-CNN through shared convolutional features. As such, Faster R-CNN can be seen as consisting of two distinct modules. The RPN network tells the unified network where to look and the Fast R-CNN detector uses the proposed regions to make predictions. Faster R-CNN achieved a mAP of 67%,

improving over its predecessor by 2% after being trained on the public VGG-16 model.

III. Dataset

This project uses the VinDr-CXR open dataset [5] of chest X-rays with radiologist annotations, collected by researchers at the Vingroup Big Data Institute from the Hospital 108 (H108) and the Hanoi Medical University Hospital (HMUH), two of the largest hospitals in Vietnam (paper). This dataset consists of 18,000 postero-anterior view chest X-ray scans that come with both the localization of critical findings and the classification of 14 common thoracic diseases. These images were annotated by a group of 17 radiologists with at least 8 years of experience for the presence of the diseases found in the dataset. The Vingroup researchers divided the dataset into two parts: the training set of 15,000 scans and the test set of 3,000 scans. Each image in the training set was independently labeled by 3 radiologists, while the annotation of images in the test set were obtained through the consensus findings of 5 radiologists. The Vingroup researchers provided labels for the training set in the form Comma-Separated Values (CSV) file format. Each entry in this file provides information about the ID of the radiologist who made a prediction, name of the image, class ID and coordinates specifying the position of the bounding box. Since each X-ray in the training set was annotated independently, consensus was not reached among radiologists and disagreements occurred. Images in the test set were labeled by five radiologists in a two-stage process. Each image was first annotated by three doctors independently, followed by two other doctors with higher levels of expertise. Doctors reviewed the findings and communicated with each other to reach a consensus about the labels in the image. Therefore, it should be noted that images in the training set might have incorrect labels, and that images in training and testing sets were labeled using different approaches.

IV. Methodology

A. Exploratory Analysis

We first analyzed the Class and Radiologist IDs data. The distribution of abnormalities and radiologists, who labeled the images, is presented in Figure 1. As seen from the figure, certain radiologists contributed far more than the others. In addition, the training set has unequal class distribution.

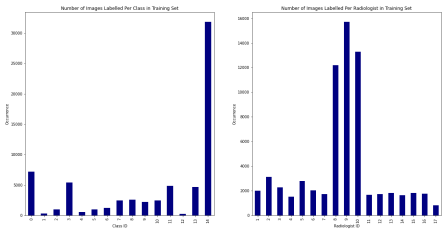

Fig. 1. Number of images per class and number of images per radiologist

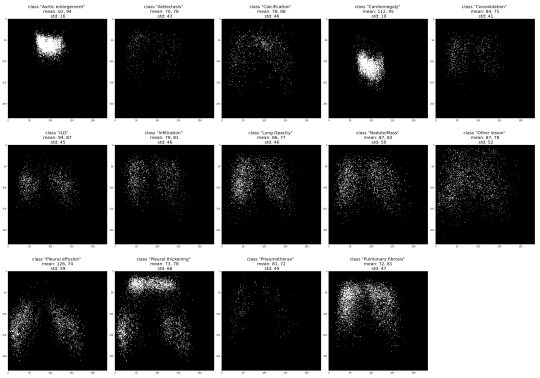

Fig. 2. Distribution of bounding box locations

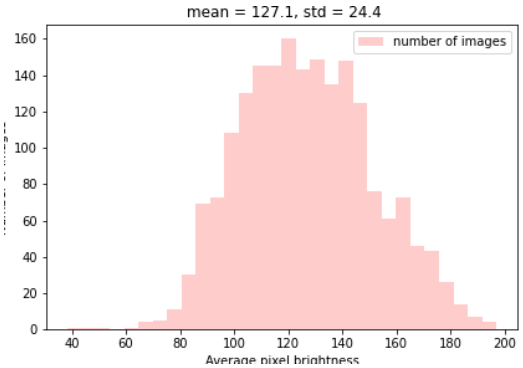

Fig. 3. Distribution of average image brightness

In order to discover if some classes tend to appear at the same spot in image, the distribution of locations for each class was plotted. Firstly, the center of bounding box was calculated for each entry in the CSV file and then it was scaled to accommodate the differences in images' sizes. Secondly, the locations of the center for each class were accumulated. And lastly, the distributions were plotted and their corresponding means and standard deviations were calculated. The distributions for each class are presented in Figure 2.

As can be seen from the figure, classes Aortic enlargement and Cardiomegaly tend to appear roughly at the same spot. In addition, compared to other classes they have the lowest standard deviations. The reason for this is that according to the nature of diseases, the first finding has to be located in the upper part of Aorta while the second in the heart area. Further, classes Consolidation and ILD tend to appear in the bronchi part of an image. In contrast, the possible positions of other findings are rather sparse: class Other Lesion is evenly distributed across the image and the rest of the classes are primarily distributed in the lung area. Thus, if a trained model has a lot of false positives for several mentioned classes, they can be easily filtered out by their location.

We also decided to analyze the pixelation and brightness of images in the dataset. First of all, the distribution of average brightness of images was plotted. This was done by scaling the value of image pixels using the maximum value that can

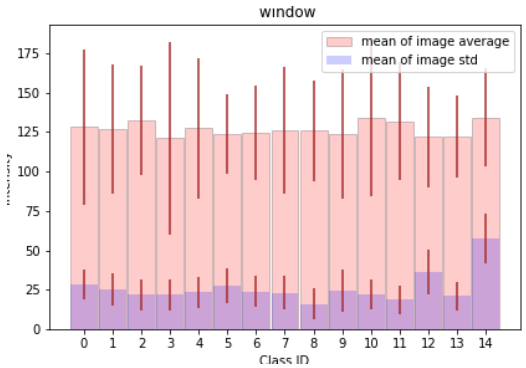

Fig. 4. Distribution of image means and standard deviations per class window.

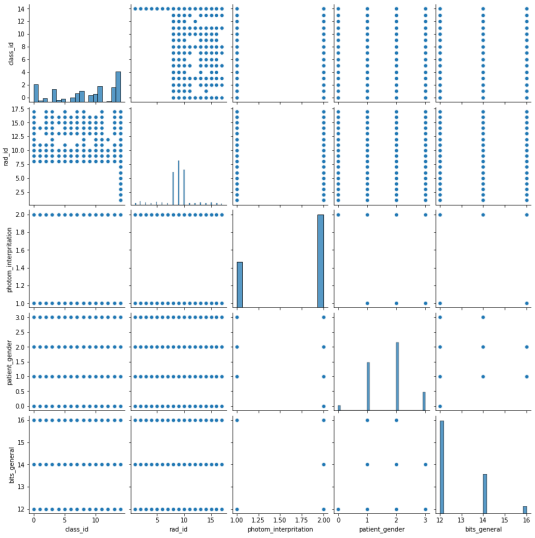

Fig. 5. Pairwise plot of image metadata: image type, patient gender, image bit depth, radiologist findings, and radiologist ID.

be stored in pixels, i.e. dividing image by two to the power of image bit depth. The mean pixel intensity was also calculated for each image. The distribution of average image brightness is presented in Figure 3.

It can be concluded that the mean brightness of images is located around the intensity of 127. However, the standard deviation is rather high, meaning that some images in the dataset are very bright or very dark. Thus, it becomes necessary to center the intensity of all images around the same value.

Next, the distribution of mean standard deviation of intensity inside the bounding boxes of each class was plotted. This was done in order to discover if some classes tend to appear in bright or dark regions of image, and how much the intensity varies on average for each class. In this way, the mean of pixels' values was calculated for each class separately. The same was done with the standard deviation of intensity inside each window. Since the last class has the full image associated with it, the mentioned values were calculated for the entire images. The resulting distribution is presented in Figure 4 Therefore, based on this distribution it can be concluded that image brightness inside the bounding boxes is roughly the

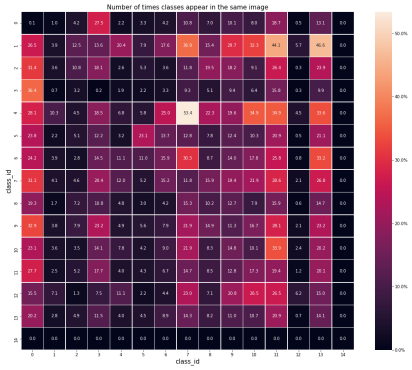

Fig. 6. Number of times classes appear in together in the same image.

same across the classes. The same applies to image standard deviations.

As a next step of data analysis, some metadata taken from the images and data from the CSV file were grouped together. The correlation table was created together with pairwise graphs (Figure 5). This allows us to see if there are any correlations between the features. As can be seen in Figure 5, there are no meaningful correlations between the data, except a slight correlation between the radiologist and class IDs, and image type and depth. Indeed, as seen from the pairwise plot, the radiologists who have IDs from 1 to 7 predict only No finding class, they have never predicted any other class. This also will be taken into account during the preprocessing step. The correlation between image type and depth can be explained by the fact that different imaging devices were used that had different presets.

Given that several different classes might appear in the same image, it is worth analysing which classes appear together. To achieve this, Figure 6 plots the number of times two classes appear together in image, divided by the total number of instances of a row class. For example, class 1 appears together with class 13 46% of the time. It can be concluded that a majority of classes may show up together in an image. Also, as can be seen from the table, class 14 which corresponds to the No finding class, never appears together with other classes, meaning that radiologists are always certain that there are no findings.

B. Preprocessing

Our data preprocessing involved three main steps. We first converted all images from DICOM format to a multidimensional array. By converting DICOM's to arrays and mean-centering all values around zero, we produce image data that is both normalized and ready to be used as input into our model. Second, we applied image augmentation to all training images with the albumentations python package. Specifically, our transformations include flipping images along the X-axis, rotating images between 0 and 20 degrees, and rescaling. We

TABLE I
HYPERPARAMETERS AND VALUES.

Hyperparameter	Values
Number of epochs	[1,2]
Learning rate	[0.001, 0.005, 0.01]
Momentum	[0.9, 0.95]
Starting weights	[Pre-trained, random]

TABLE II
HYPERPARAMETER TUNING RESULTS.

Model	Hyperparameters				Loss Value
	Epoch	Learning Rate	Momentum	Starting Weights	
ResNet50-FPN	1	0.005	0.95	Random	0.153
MobileNetV3Large320p	1	0.01	0.9	Pre-trained	0.108
MobileNetV3Large180p	1	0.001	0.95	Pre-trained	0.181

converted all relevant data including image arrays, bounding box data, radiologist findings, and radiologist IDs into PyTorch tensors to feed into our model.

C. Models

As discussed in the related work, R-CNN models are an ideal choice for object detection with localization tasks. So far, our has used the following three Faster R-CNN models: ResNet50-FPN, high resolution MobileNetV3Large and low resolution MobileNetV3Large.

D. Hyperparameter Tuning

In total four hyperparameters were tuned for each of the Faster R-CNN models tested. The hyperparameters together with their tested values are presented in Table I. The best hyperparameters for each of the tested models together with their loss values are presented in Table II. So far, the best performing model is *MobileNetV3Large320p*. Its training and validation losses are presented in Figure 7.

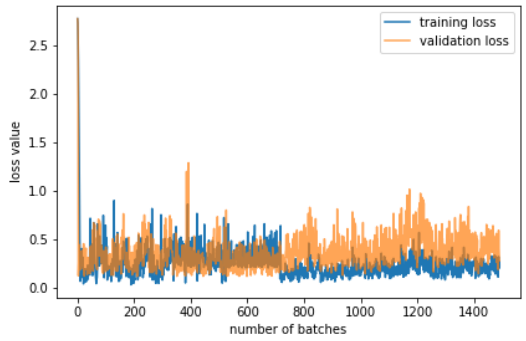

Fig. 7. Training and validation loss for the optimal hyperparameters of the Faster RCNN MobileNetV3 Large model.

V. RESULTS AND ANALYSIS

A. Metrics

From the literature we surveyed, the default and most popular metric used to measure the accuracy of Faster R-CNN object detectors is mean average precision (mAP). mAP computes the average intersection over union thresholds (IoU) for the set of predicted bounding boxes and ground truth boxes across images in a dataset. Formally, the IoU for ground truth box A and predicted box B is defined as follows:

$$IoU(A, B) = \frac{A \cap B}{A \cup B}$$

In the literature, the most common threshold to satisfy intersection has been 0.5. In other words, a predicted object is considered correct if its intersection over union with a ground truth object is greater than or equal to 0.5. For each pair of boxes, a precision value is calculated as:

$$\frac{TP}{TP + FP + FN}$$

mAP is thus the averaged sum of all precision values over a dataset.

B. Benchmarking Expected Results

As of March 21 there are 1159 teams participating in the competition. The top five teams have an average mAP of 0.3374%. Currently, our best performing model has a mAP of 0.137%. We expect our score to increase as we perform more extensive hyperparameter tuning and experiment with image augmentations.

VI. NEXT STEPS

One of our concerns is the amount of disagreements between radiologist findings, as described in Section IV. We plan on training our models on subsets of the dataset that minimize the amount of conflicting ground truth labels. We also plan on inspecting where our model is performing poorly, and experimenting with strategies to address those weaknesses.

REFERENCES

[1] A. Rimmer, "Radiologist shortage leaves patient care at risk, warns royal college," *BMJ: British Medical Journal (Online)*, vol. 359, 2017.

[2] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[3] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *arXiv preprint arXiv:1506.01497*, 2015.

[5] H. Q. Nguyen, K. Lam, L. T. Le, H. H. Pham, D. Q. Tran, D. B. Nguyen, D. D. Le, C. M. Pham, H. T. Tong, D. H. Dinh *et al.*, "Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations," *arXiv preprint arXiv:2012.15029*, 2020.