Diagnosing and Locating Abnormalities in Chest X-Rays using Deep Learning

G032 (s2117072, s1314602, s2125570)

Abstract

Chest X-Rays are one of the most common screening procedures used to diagnose a range of medical conditions including tuberculosis and lung cancer which affect millions of people every year. Advances in DL have led to the development of computer aided diagnosis (CAD) of medical images, achieving performance matching expert radiologists. While previous chest X-Ray analysis systems are able to diagnose abnormalities, they cannot show their location on the image. Chest radiographs are also some of the most challenging to interpret, and are typically analysed by a team of radiologists to come to a consensus. The release of a new high quality dataset from the Vingroup Big Data Institute containing annotated X-Rays with the locations of abnormalities now makes it possible to develop a system with this capability. In this paper we train a state of the art object detection model on this new dataset and improve our baseline validation AP50 score from 0.280 to 0.593 through our experimentation. Our final submission generates a Kaggle leaderboard score of 0.204. These results are also reviewed against other state of the art models in the literature.

1. Introduction

The World Health Organization estimates that two-thirds of the world population does not have access to diagnostic medical imaging, driven by a shortage of expert radiographers (Welling & et al., 2011). Chest radiography is one of the most common procedures with over 2 billion examinations performed every year. It is used to screen and diagnose several conditions including cardiothoracic and pulmonary abnormalities which are among the leading causes of mortality and health service use in the world (Mathers & Loncar, 2006). Research in computer-aided diagnosis (CAD) where computer outputs are assumed to aid radiologists in detecting and diagnosing health conditions has started in the 1960s and from the 1990s this has been one of the major subjects of research in medical imaging (Doi, 2007; Giger & Suzuki, 2008). This technology has had a significant impact in medicine, for instance, breast cancer rate reduced by over 30 percent in United States since mammography got widespread in 1990 (Brook, 2012). With an exponential increase of interest in deep learning (DL) and significant advancements in performance of DL algorithms over the last decade, DL models were able to reach expert-level performance. For example, model trained to classify interstitial lung diseases in chest X-Ray images achieved 85.5% accuracy (Anthimopoulos et al., 2016).

There are several key advantages that DL based systems bring to the area of radiology. First of all, these systems address the problems stemming from human factor. Overwhelmed with work radiologists might delay reviewing X-Ray scans and not detect potentially life-threatening conditions in time or make wrong conclusions (Newell & Rosenbloom, 1981). By combining output from the model and radiologist knowledge, more accurate diagnosis can be made. In addition to that, automated radiology systems can increase the access to radiology especially in the regions where health services are limited. Moreover, machine learning algorithms have demonstrated the ability to discover patterns in chest X-Ray scans that are generally not noticed by doctors such as slightly varied pose of of patients with (Reardon, 2019). Finally, models trained on images annotated by multiple radiologists can be more accurate than single radiologist and overcome the bias (Mitchell, 1980).

Despite these advancements in medical imaging and AI, there are some important challenges related to use of machine learning in radiology. While there are many publicly available datasets of chest X-Ray scans such as ChestX-Ray14, ChestX-Ray8 (Wang et al., 2017), Padchest (Bustos et al., 2020), CheXpert (Irvin et al., 2019), and MIMIC-CXR (Johnson & Pollard, 2019), they are annotated using text mining and Natural Language Processing techniques to parse the medical reports associated with each X-Ray, and the accuracy of these annotations poses significant issues (e.g. inconsistency, uncertainty, errors) that affect the quality of the DL models trained with them (Oakden-Rayner, 2017). Additionally, the locations of the abnormalities are not present in these datasets. Although annotating X-Ray images requires significant human resources, it would benefit the development of models. Majority of commercially used models are trained on private data and should the system make wrong decision, it may be difficult to explain what has caused certain outcome (Reardon, 2019). This issue relates to ethical dilemma of determining the liability of wrong diagnosis if doctor's decision was based on AI system (Reardon, 2019). U.S. Food and Drug Agency (FDA) has not yet verified standards regulations for machine learning applications in medical imaging that would give more clarity to the field (Kohli et al., 2017) as suggested in (Thrall et al., 2018). Finally, to the contrary of experiments in (McBee et al., 2018), it is believed that

models trained on X-Ray scans from particular institution would not be able to function appropriately given samples from other systems.

To address these issues and accelerate research in machine learning applications within area of radiology, the Vingroup Big Data Institute (VinBigData) released a new dataset, VinDr-CXR (Vin-Dr), the largest public dataset of chest X-Rays annotated by expert radiologists and containing the locations of the diagnosed abnormalities (Nguyen et al., 2021). In general, areas of active research in DL applications within radiology can be grouped into four overlapping categories: lesion and disease detection, classification and diagnosis, segmentation and quantification (McBee et al., 2018). In this paper we explore object detection and classification problem. With this context and challenges in mind, we aim to answer the following research questions:

1. *Can we can build a model that makes accurate predictions for abnormality detection in chest X-Rays?*
2. *Can we address imbalances in the abnormality classes in the dataset and handle noisy bounding box annotations from multiple radiologists to improve model performance?*

To achieve our goal we train two modern object-detection models on the VinDr dataset and analyse their performance.

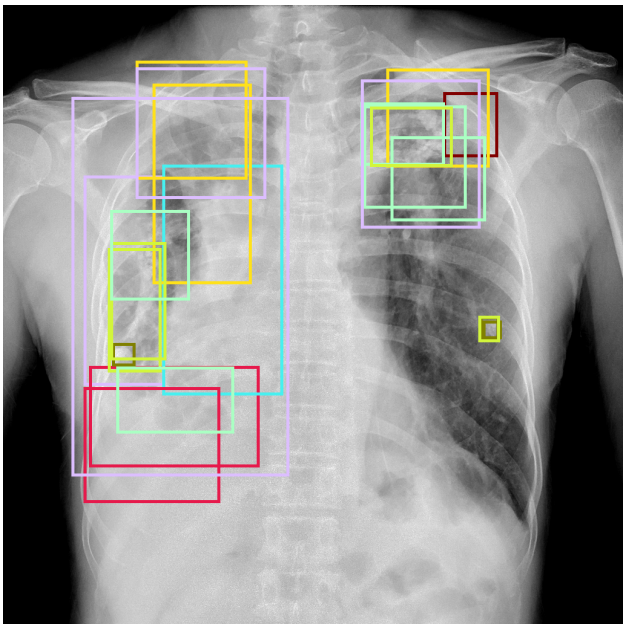

Figure 1. Example image from the VinBig dataset with drawn ground-truth annotations by radiologists.

2. Dataset

To conduct our experiments, we used the newly released dataset VinDr by VinBigData that is also part of the "VinBigData Chest X-Ray Abnormalities Detection" Kaggle competition (Institute, 2020). It was released with a report (Ha Q. Nguyen, 2021) detailing how the data was gathered, processed and annotated. The full dataset contains over 100,000 chest X-Ray images collected from two major hospitals in Vietnam from 2018 to 2020. Due to a significant portion of the data containing low quality images, personal identifiable information or scans of other body parts, a random sample of 18,000 images were available for the competition. 15,000 of these images had available annotations (training set) whilst 3,000 of chest X-Ray images were used to benchmark competition participants and thus their labels were private (test set). There are a total of 67914 available annotations. All images were in DICOM, a standard format in medical imaging (Association) and so required preprocessing as discussed in Section 2.2. See Figure 1 for examples of X-Ray images and their associated annotations.

One of the main features of this dataset making it distinct from other chest scan collections are its hand-labelled annotations. A team of 17 highly-experienced radiologists annotated images in both the training and validation set with 3 radiologists assigned to annotating each training image, resulting in three sets of annotations per images. However the test set, contains a single set of annotations that have been generated through a consensus of 5 radiologists. All the data was de-identified and manually reviewed according to data protection laws thus preserving patients privacy.

The key task presented for this dataset is to predict the abnormality and its associated bounding box identified by the team of radiologists.

2.1. Exploratory data analysis

There are 15 possible classes for each annotation with 14 classes identifying different abnormalities whilst the last class identifies no abnormalities present. In diagrams and charts we use consistent color scheme to identify different abnormality classes (by name). We carry out some exploratory data analysis (EDA) to gain insights into the data and found the following:

- Significant class imbalance: 70.7% of images do not contain any abnormality, thus there are only 4394 images to train the model to detect abnormalities.
- Majority of images in abnormality classes contain more than one annotation per image (can be of multiple classes).
- Noisy annotations: there can be multiple annotations in image that are meant to represent exact same instance of abnormality, thus there are more annotations in total than available images.

The number of images containing each abnormality as well as total number of annotations for each class is shown in Figure 4. Although there are a large number of training images in the dataset there is a clear class imbalance. Our models are likely to struggle to predict locations of abnormalities within these underrepresented classes, especially for conditions such as Atelectasis or Pneumotrax for which

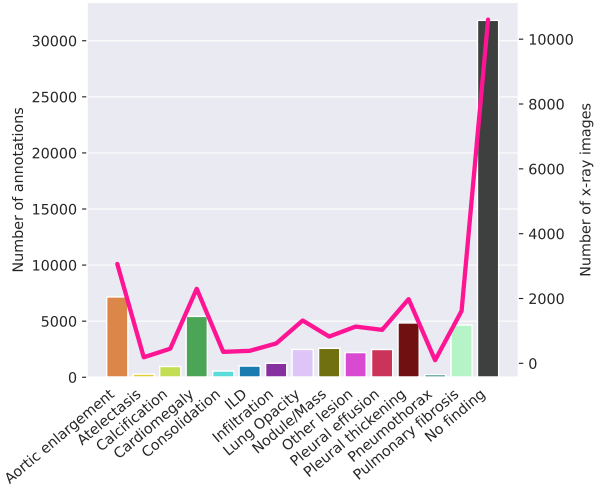

Figure 2. Distribution of annotations and images in the training set original VinDr dataset.

there are only 186 and 96 images containing in total respectively. Imbalances within the dataset can generate class biases in the model where abnormalities that are more common in the training data will be more likely to be predicted. This can become a significant problem if the balance of the classes changes in future data as this could lead to poor generalisation performance.

2.2. Preparing data for experiments

To conduct our experiments, we split the original training data into training, validation and test sets with a ratio of 80/10/10 for each split. When discussing the results further in this paper, we refer to data in this split. Bounding boxes for 'No abnormaility' are set to (0, 0) (1, 1) as per the Kaggle submission guidelines.

2.3. Generating average and supremum annotations

Each X-Ray is annotated by 3 radiologists which leads to multiple annotations for each abnormality. The original test set of 3,000 images only contained a single annotated box per image, however since we split our test set from the training set, we want to consolidate the bounding boxes so that there is one box per abnormality, matching the original dataset. The theory is that consolidating the bounding boxes in the training data will make it easier for the model to make predictions, as it only has to correctly predict one bounding box per abnormality, rather than two or three.

We devised two strategies for consolidating bounding boxes which we call 'average' and 'supremum'. We process the annotated boxes for each image by first grouping together boxes for the same abnormality class. For any two boxes of the same class that overlap we either take the average of the bounding box co-ordinates to get the 'average' box, or in the second strategy take the smallest bounding box that contains both boxes to get the 'supremum' box. The

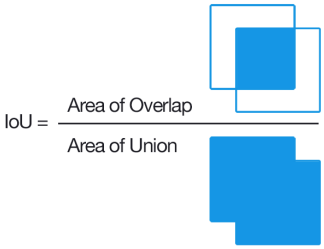

Figure 3. Visualisation of intersection over union metric.

algorithm is repeated recursively until the bounding boxes within each abnormality class are disjoint. The effect on the number of, location, and size of the annotated boxes can be seen in Figures 4 and 5. Observe that the distribution of number of bounding boxes of each class doesn't change significantly, indicating that the radiologists predictions are mostly in consensus (mostly overlapping). Observe that the widths and heights of the supremum boxes are more distributed with more boxes covering larger portions of the image. We will investigate which of these three strategies produces a better model.

2.4. Performance metrics

We primarily evaluate the performance of our models using the Average Prevision 50 score (AP50). To explain this metric we must first identify what is classed as a correctly predicted bounded box. A correct prediction is when the intersection over union (IoU) of the predicted and target bounding boxes is over 0.50, see Figure 3 for a visualisation of this. We then use the precision - as shown in equation 1 where tp and fp is the true positive and false positive counts respectively - and averages over all data points.

$$precision = \frac{tp}{tp + fp} \tag{1}$$

3. Methodology

As described in Section 2 we evaluate two model configurations. The first configuration is an object detection model trained on all images, including negative samples, in the training set. The second is an ensemble of a binary classifier to distinguish X-Rays containing abnormalities from those that don't and an object detection model trained only on positive samples (X-Rays that contain abnormalities). The models are trained using the methods for consolidating annotations and addressing the class imbalance described in section 2.

3.1. Detectron2 vs Pytorch

We initially started our approach to this task using the Detectron 2 framework (Wu et al., 2019). Detectron 2 is a platform for object detection and segmentation built by

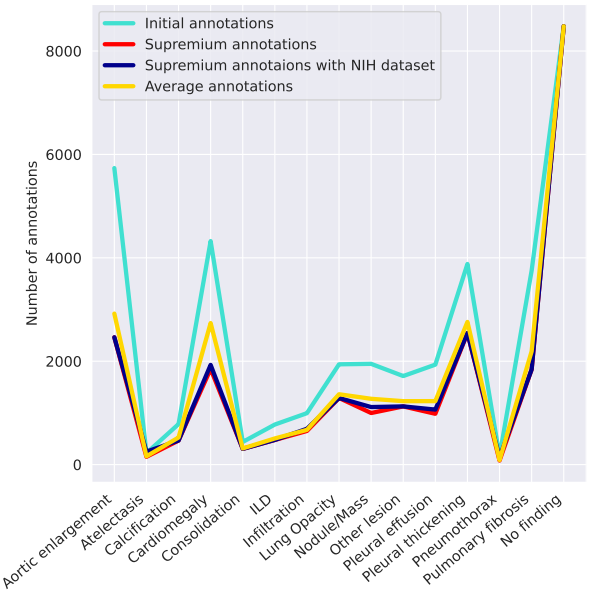

Figure 4. Distribution of annotations and images after applying different processing techniques for annotations in set used for training.

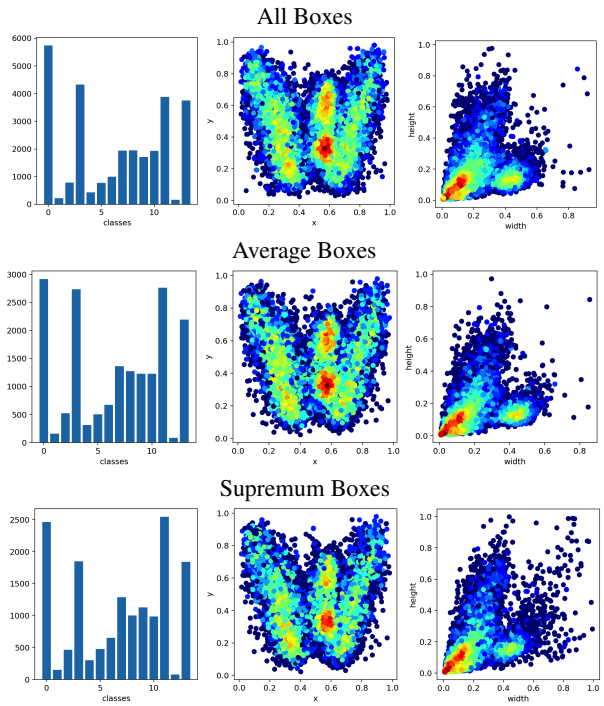

Figure 5. Charts describing the annotated boxes in the training set: 1. Bar chart of the number of annotated boxes for each abnormality class. 2. Distribution of the normalised center of the annotated boxes. 3. Distribution of the normalised width and height of annotated boxes

Facebook AI Research (FAIR) that provides state-of-the-art models. From our experiments we found this framework to be very complex and required us to overwrite many classes to give us the flexibility we needed. The models we used such as YOLOv4 and Faster R-CNN were available in the Pytorch framework whilst also providing easier implementation and more flexibility. All models used in our experiments were done so in the Pytorch framework.

3.2. Faster R-CNN

In contrast to YOLOv4 as described in Section 3.3, Faster R-CNN is a two stage system that is comprised of Region Proposal Network (RPN) and Fast R-CNN modules (Ren et al., 2016). The RPN is a fully convolutional network that generates regions of interest. These proposals are then passed to the Fast R-CNN module which detects objects within these regions. One of the significant contributions of the Faster R-CNN architecture is the shared computation between these two modules. As can be seen in Figure 6 the final convolutional layer is used both modules which improves computational efficiency. This architectural design choice also allows both modules to be trained in a single network which was not possible in previous designs (Girshick et al., 2014)(Girshick, 2015).

Faster R-CNN has generated state-of-the-art results and have been the foundation of systems that won COCO 2015 competitions and generates AP.50 scores of 58% (Ren et al., 2016).

3.2.1. Augmentations

Image augmentation is an effective technique for improving the performance of computer vision (CV) systems. It is often difficult to obtain sufficient data for a CV task, but image augmentation can mitigate this issue.

Image augmentation is a technique of modifying images in such a way that can be thought of generating new training data by slightly changing the original images. Image augmentation has been shown to improve model performance across a range of architectures including AlexNet, ResNet and EfficientNet (Krizhevsky et al., 2012)(He et al., 2015)(Tan & Le, 2020). In our experiments a set of augmentations were applied to various models, these augmentations consisted of random croppping, horizontal flip, brightness and contrast augmentations. Through the use of the Albmentations library these augmentations modify the bounding box in the appropriate way. For example, a random cropping could remove some of the bounding box, however, a new bounding box would be generated to account for this transformation (Buslaev et al., 2020).

3.2.2. Weight Balancing

As discussed in the Section 2 there is a severe class imbalance between the different abnormalities. As part of the Faster R-CNN architecture the final layer uses a cross-entropy loss function which we can manually rescale to account for the class imbalance. The class weights were calculated using Equation 2 to linearly rescale them, where α is class count, w_k is the weight for class k, and n is

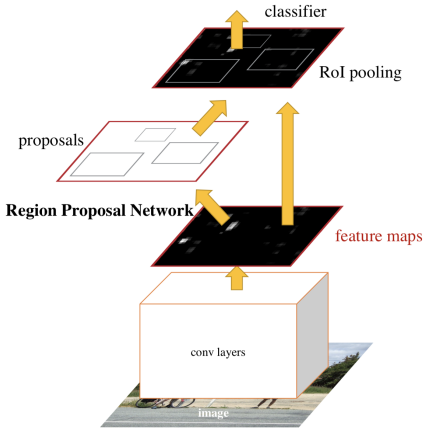

Figure 6. Faster R-CNN shares computation between the Region Proposal Network and ROI pooling to improve computational efficiency and accuracy. Figure is taken from Faster R-CNN paper (Ren et al., 2016)

the number of classes. This approach also calculates the weights so they average to 1 to prevent any changes in "effective" learning rate to help experiments be directly comparable. The loss for the new weighted cross entropy error can be seen in Equation 3.

$$w_k = \frac{n}{\alpha_k \sum_{i=1}^{n} \frac{1}{\alpha_i}} \tag{2}$$

$$loss(x, y) = w_k(y)(-x(y) + log(\sum_j \exp(x_j))) \tag{3}$$

3.3. Scaled YOLOv4

Scaled YOLOv4 (You Only Look Once) is a state of the art object detection model, achieving between 66% and 73% AP.50 scores on COCO depending on the configuration (Wang et al., 2021). The model architecture uses a one stage architecture which detects objects in the image by scanning it only once. We use the Scaled Yolov4-CSP which can be quickly trained on a single GPU, the other configurations, p4, p5, and p7, requiring multiple GPUs to train which was outside the hardware capabilities available to us. We employ transfer learning, starting with pretrained weights which achieve 66% AP.50 on COCO.

To combat the class imbalance in the training set the class losses are scaled by the normalized class weights (Phan & Yamamoto, 2020). YOLO computes the class losses using binary cross entropy loss with logits which provides two options for scaling the losses: setting normalised class weights, or setting a 'pos_weight' that scales the class loss scores by the ratio of negative to positive training samples in that class. In preliminary experiments setting the 'pos_weight' performed worse than setting normalised class weights so we implemented the latter approach. The normalised class weights $\hat{w_k}$ are computed from the total number of classes N and the number of annotations in class k n_k, using formula 4.

$$\hat{w_k} = \frac{N \cdot w_k}{n_k \cdot \sum_i 1/n_i} \tag{4}$$

3.4. Binary Classifier - Inceptionv3

The task described in Section 2 is very complex and typically produced much lower performance scores compared to other tasks such as COCO. With an aim to simplify this task we introduce a binary classifier to identify if an image has an abnormality or not. Following this Faster R-CNN will be applied to identify the specific abnormality if required. The motivation behind this approach is that a state-of-the-art specialised classifier is likely to perform better than the classifier in the Faster R-CNN architecture. This approach also breaks the task into two steps which we are expecting to be easier than a single model approach. A further benefit of this approach is that we can implement class balancing methods such as oversampling to reduce bias in our model. This is possible because the ratio of abnormal to non-abnormal datapoints is roughly 2:1 and the complexity of multi class labels no longer exist in our dataset - trying to oversample multi class labels is a complex problem especially when trying to smoothly sample across classes (see future work in Section 6 for more detail).

The various inception architectures are based upon the principles of creating wider rather than deeper networks in order to mitigate issues such as vanishing gradients, overfitting and computational expense that are common in deep networks (Szegedy et al., 2014). By applying various kernel sizes on the same layer, the inception architectures generate wide networks that can capture global and local trends from the input layer. These filters are concatenated together before being passed to the next layer (Szegedy et al., 2015).

Inception v3 is the third generation of this architecture that incorporates a number of upgrades that reduce computational complexity and increase accuracy. Firstly convolutions can be factorised; an n×n convolution can be performed by a 1×n and n×1 convolution resulting in significant reduction in computational expense. Inception v3 also uses rigorous regularizing including label smoothing and batch normalisation in the auxiliary classifiers - auxillary classifiers are classifier heads that are included in layers before the final layer to help convergence of the network.

Inception v3 has shown near state-of-the-art results with top 1% accuracy of 78.8% on ImageNet (Szegedy et al., 2015)

4. Experiments

4.1. Object Detection on All Training Images

Following the methodology described in section 3 we obtained a set of baseline models trained on all the images and annotations in the training set including the negative samples that don't contain any abnormalities. Faster-RCNN was trained for 25 epochs with the Adam learning rule with

learning rate 0.0001 and no regularization. Scaled-Yolov4 was trained for 40 epochs using SGD with momentum with all hyperparameters set to default values: learning rate 0.01 and momentum 0.937. Faster-RCNN and YOLOv4 are already pre-configured with suitable default values for the hyperparamters and do not require much tweaking. Only the learning rate for Faster-RCNN had to be changed from the default to avoid exploding gradients and no additional regularization was required as the validation scores stabilized and showed no signs of overfitting. The results are displayed in the first row of Table 1. The AP.50 scores are low, all less than 30% indicating the difficulty of task. We find that YOLOv4 outperforms Faster-RCNN by 10%, a significant margin which is likely down to the more advance model architecture.

4.2. Classification and Object detection

The first component of the ensemble model configuration is a binary classifier to distinguish between abnormalities and no abnormalities. Inceptionv3 was trained on all the images in the training set, oversampling the non abnormality images to balance the two classes. The model was trained for 25 epochs with learning rate 0.0001 attaining 80% accuracy on the validation set.

For the second component of the ensemble we train object detection models using only the training images that contain abnormalities. As described in section 2 we train the models using three sets of annotated bounding boxes: 'All' uses the training boxes from all three radiologists, 'Sup' takes the supremum bounding box of any pair of overlapping boxes identifying the same abnormality, 'Avg' takes the average of overlapping bounding boxes identifying the same abnormality. By training on all three datasets we can see which consolidation method produces the best model.

For each of the three sets of annotations, faster-RCNN was trained for 25 epochs using learning rate 0.0001. The results of the experiments are summarized in Table 1. The model performed best using the supremum boxes with an AP.50 of 0.31% on the validation set, significantly higher than 0.25 on the average boxes and 0.24 when using all boxes. Scaled-YOLOv4 was trained for 40 epochs on the same sets of boxes with default hyperparameters producing the same trend with the best AP.50 score of 0.35 on the supremum boxes and a lower score of 0.31 on the other two sets of boxes.

The large difference in performance between the supremum boxes and the average and all boxes is surprising. As explored in Section 2 the distribution of 'supremum' boxes isn't significantly different from the 'average' boxes, the only notable difference being the increased area. Additionally the 'average' boxes did not perform much better than with all boxes present. One theory that explains this is the 'average' boxes do not accurately capture all the abnormality - sometimes the box is off center and a portion of the abnormality is left outside the box.

4.3. Class Imbalance

Finally we investigated the effect of adjusting the class losses on model performance as described in Sections 3.3 and 3.2.2. For each of Faster-RCNN and YOLOv4, two models were trained using all the annotated boxes, one with adjusted losses and the other without, and a breakdown of the AP.50 scores per abnormality class on the validation set were computed, see Table 2. The effect on performance is mixed. Across all abnormalities, the models without weight adjustment had higher AP.50 scores, a 1% improvement in the case of YOLO. One explanation for this is the validation set has almost the same distribution of abnormalities as the training set, so if the model is biased towards certain abnormality classes this will translate into better validation performance.

Looking at the YOLO scores on individual abnormality classes, some poorly represented abnormality classes such as Pneumothorax and Consolidation performed better with weight adjustment, whereas others including Atelectasis and ILD performed worse. From the theory (section 3) we expect validation performance to improve on poorly represented abnormality classes when compensating for the class imbalance, but adjusting the losses during training has had mixed results. We did not investigate exactly why some classes improve while others perform worse, but we speculate this could be down to interference between the loss scaling and the other data imbalance corrections that YOLO employs such as focal loss, which attempts to address the imbalance between foreground and background classes. On the other hand Faster-RCNN performed better across all types of abnormality with weighted loss than without. This is likely due to the lack of any other data imbalance methods in Faster-RCNN, so doing any kind of compensation for class imbalance would likely yield an improvement across all classes.

4.4. Ensemble of Classifier and Abnormality Annotators

Another idea we explored was to construct an ensemble of a single classifier model that determines which abnormalities are present in an X-Ray followed by a collection of annotation models, one for each abnormality class, that predict the locations of the specific abnormality. The intuition is that specific annotation models will have better performance than a single model that is capable of annotating every type of abnormality as it's a much simpler task for the model to learn. Additionally there exist other chest X-Ray datasets such as ChestX-Ray8 that contain images labeled with the types of abnormalities present, but not their locations (Wang et al., 2017). These could be used to balance out the dataset for the initial classifier. The annotator models do not need mitigations for class imbalance as they are only trained on one type of abnormality. However the performance of the individual annotator models was very poor due to the lack of training data, and due to lack of available GPU time required to train all the models this experiment was dropped.

4.5. Model Selection and Evaluation

Scaled-YOLOv4 outperformed Faster-RCNN across all experiments. The best performing system was Scaled YOLOv4 with the supremum annotated boxes. We configured our ensemble model to first feed images through the binary classifier and, if it determined abnormalities were present, fed the images into the object detection model. We evaluated our model on the test set using the supremum annotated images yielding an AP.50 score of 0.582. The score is high compared to the scores for the object detection models because most of the images do not contain abnormalities and the binary classifier had a high accuracy of almost 80%.

We also submitted our best model to the Kaggle competition, evaluating on the original VinDr test set and obtained a private AP.40 score of 0.204, compared with the winning score of 0.314.

5. Related Work

The advances of DL in the fields of image classification and object detection have led to increasing interest in applying the technologies to medical image diagnosis, particularly chest X-Rays which is the most common imaging procedure. As stated in the introduction of this paper DL requires a large amount of training data to produce effective models. However existing datasets of chest X-Rays do not contain sufficient annotated images, due to a combination of patient privacy regulation preventing the release of the X-Rays and the high cost of getting expert radiologists to annotate the X-Rays. While in this paper we employed data augmentation and loss scaling to combat the imbalances, researchers have applied other novel methods to combat these issues.

Due to there being no large publicly annotated chest X-Ray dataset that contained bounding boxes for the locations of abnormalities before the VinDr dataset there are no existing published results for this type of task we can directly compare against. However, we did evaluate our best model on the Kaggle test set and submit our score for ranking. Our results were average for the competition, achieving 0.204 AP.40, with the winner of the competition achieving 0.314. The winning system used an ensemble of models to produce an overall prediction, and then adjusted the classification probabilities of the type of abnormality present in each predicted box to achieve a high score[1].

The closest X-Ray abnormality detection task is classifying abnormalities present in images without finding their locations. (Wu et al., 2020) perform a survey of chest X-Ray diagnosis techniques and compare their performance against human experts. The conclusion from the research is that while DL models can match human performance for simple diagnosis, more complex cases are better handled by humans.

[1]https://www.kaggle.com/c/vinbigdata-chest-xray-abnormalities-detection/discussion/229724

One novel approach to address the lack of training samples for rare chest abnormalities is to use Generative Adversarial Networks to create synthetic X-Rays. (Kora Venu & Ravula, 2021) and (Madani et al., 2018) generate synthetic images of under represented samples and show an improvement in the models trained on the extended datasets, even when the original training set contained around 1300 images.

The current largest labeled Chest X-Ray dataset, ChestX-Ray8 contains over 100,000 images that were labeled using NLP techniqes to parse the associated medical reports for the abnormalities present in the images, for which the authors estimate is over 90% accurate (Wang et al., 2017). A recent novel approach to automate the annotation of bounding boxes for chest X-Rays uses a dataset of radiologists dictation audio and eye gaze co-ordinates to train a model capable of this task and demonstrate its feasibility (Karargyris et al., 2021). Another major issue with annotations is the variability in the quality of labels when X-Rays are annotated by independent radiologists. If radiologists use a majority vote approach to determine labels then this could filter out important but difficult to find abnormalities that were only identified by a small number of radiologists. Excluding these labels can make a significant different to the performance of models trained on the data. (Majkowska et al., 2020) investigate this issue and employed a team of radiologists to verify the labeling of subsets of their dataset, taken from two existing dataset ChestX-Ray 14 and DS1, and used 'population adjustment' to evaluate the performance of their models to take class imbalances into account.

6. Conclusions

We've demonstrated that it is possible to train a DL model to diagnose and locate abnormalities in chest X-Rays. We identified and explored two major issues with chest X-Ray datasets: the lack of training images, especially for rare abnormalities leading to high class imbalance, and consolidating noisy annotations from multiple radiologists. We devised approaches to solve both of these issues and ran experiments using two state of the art object detection models, Faster-RCNN and Scaled-YOLOv4, to compare their effectiveness.

To combat class imbalance we applied augmentations to the images to generate more data points, and adjusted the class losses proportionally based on the number of annotations in each class. This had a positive effect for Faster-RCNN improving average precision scores across all classes, but had mixed results with YOLOv4 with some abnormality classes improving and others getting worse with an overall drop in model performance. We speculate that weight adjustment worked with Faster-RCNN as it has no built in methods for dealing with class imbalance, whereas YOLOv4 already has several built in algorithms such as focal loss to combat dataset imbalance and this may be interfering with the additional loss adjustments.

To combat noisy annotations we devised two approaches to

Model	Data	Augmentations	Loss Weighting	Annotations	Validation AP.50
Scaled-YOLOv4	All	Yes	Yes	All	0.280
Faster-RCNN	All	No	No	All	0.203
Faster-RCNN	Only Abnormalities	Yes	No	All	0.243
Faster-RCNN	Only Abnormalities	Yes	Yes	All	0.245
Faster-RCNN	Only Abnormalities	Yes	Yes	Sup	0.318
Faster-RCNN	Only Abnormalities	Yes	Yes	Avg	0.258
Scaled-YOLOv4	Only Abnormalities	Yes	No	All	0.330
Scaled-YOLOv4	Only Abnormalities	Yes	Yes	All	0.318
Scaled-YOLOv4	Only Abnormalities	Yes	Yes	Sup	0.345
Scaled-YOLOv4	Only Abnormalities	Yes	Yes	Avg	0.310

Table 1. Object detection model results.

Class	No. Annotations	Faster-RCNN	Faster-RCNN w/ Weighted Loss	YOLOv4	YOLOv4 w/ Weighted Loss
All	3702	0.243	0.245	0.330	0.320
Aortic enlargement	737	0.430	0.449	0.677	0.669
Atelectasis	29	0.227	0.370	0.335	0.273
Calcification	82	0.188	0.214	0.158	0.117
Cardiomegaly	547	0.424	0.423	0.615	0.581
Consolidation	50	0.155	0.350	0.311	0.329
ILD	93	0.243	0.456	0.379	0.356
Infiltration	129	0.205	0.338	0.270	0.299
Lung Opacity	265	0.220	0.238	0.256	0.267
Nodule/Mass	296	0.247	0.275	0.261	0.219
Other lesion	228	0.134	0.129	0.136	0.108
Pleural effusion	275	0.315	0.414	0.418	0.362
Pleural thickening	504	0.187	0.300	0.263	0.250
Pneumothorax	28	0.206	0.318	0.296	0.408
Pulmonary fibrosis	439	0.224	0.214	0.244	0.246

Table 2. Class breakdown of AP.50 scores of models trained on just the abnormalities using all the annotated boxes, comparing with and without weighting the losses to account for class imbalance.

consolidate overlapping bounding boxes of the same type of abnormality: 'average' taking the average of overlapping boxes, and 'supremum' taking the smallest bounding box that contains all the overlapping boxes. Ideally we would have got a team of expert radiologists to verify the consolidated annotations are still acceptable for medical diagnosis, but in the absence of that we analysed the distribution of the size and location of the bounding boxes which broadly remained the same for both consolidation approaches, lending confidence that they are effective as they did not drastically alter the dataset. Our experiments showed that the 'supremum' method performed best and we hypothesize this is because the boxes are better at capturing rare but essential instances of abnormalities in the X-Ray that only a minority radiologists were able to spot.

Our performance in the Kaggle competition was not very competitive and our best AP40 score of 0.204, which we didn't submit until after the deadline, would have put our result in the top 800 and was some way behind the winning AP40 score of 0.314.

There are several lines of future research that would be interesting to explore. To combat class imbalance a recently published algorithm MLSMOTE (Charte et al., 2015) can be used to oversample multi-label multi-class datasets, and it would be interesting to see if this performs better than adjusting the class losses. Additionally the ChestX-Ray8 dataset published by the National Institute of Health contains a small sample of annotated X-Rays that could be combined with the VinDr dataset and help balance it out. While we trained one model on this combined dataset which showed improved performance we did not have time to retrain all our models to compare. There also exist several other algorithms to combat multiple overlapping bounding boxes such as non-maximum suppression which are typically applied in the inference stage, but could also be applied to training labels too. Recently a new algorithm, weighted boxes fusion was published yielding better performance than existing methods (Solovyev et al., 2021). It would be interesting to compare the performance of these algorithms with the simple consolidation methods we used.

References

Anthimopoulos, Marios, Christodoulidis, Stergios, Ebner, Lukas, Christe, Andreas, and Mougiakakou, Stavroula. Lung pattern classification for interstitial lung diseases using a deep convolutional neural network. *IEEE Transactions on Medical Imaging*, 5 2016.

Association, The Medical Imaging Technology. Dicom. URL https://www.dicomstandard.org/.

Brook, Oak. International day of radiology to recognize more than a century of lives saved and improved by radiologists and medical imaging exams, 2012.

Buslaev, Alexander, Iglovikov, Vladimir I., Khvedchenya, Eugene, Parinov, Alex, Druzhinin, Mikhail, and Kalinin, Alexandr A. Albumentations: Fast and flexible image augmentations. *Information*, 11(2), 2020. ISSN 2078-2489. doi: 10.3390/info11020125. URL https://www.mdpi.com/2078-2489/11/2/125.

Bustos, Aurelia, Pertusa, Antonio, Salinas, Jose-Maria, and de la Iglesia-Vayá, Maria. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis*, 66:101797, Dec 2020. ISSN 1361-8415. doi: 10.1016/j.media.2020.101797. URL http://dx.doi.org/10.1016/j.media.2020.101797.

Charte, Francisco, Rivera, Antonio J., del Jesus, María J., and Herrera, Francisco. Mlsmote: Approaching imbalanced multilabel learning through synthetic instance generation. *Knowledge-Based Systems*, 89:385–397, 2015. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2015.07.019. URL https://www.sciencedirect.com/science/article/pii/S0950705115002737.

Doi, Kunio. Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. *Computerized Medical Imaging and Graphics*, 2007.

Giger, Maryellen L. and Suzuki, Dr. Kenji. Computer-aided diagnosis. In *Biomedical Information Technology*. academic Press, 2008.

Girshick, Ross. Fast r-cnn. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pp. 1440–1448, USA, 2015. IEEE Computer Society. ISBN 9781467383912. doi: 10.1109/ICCV.2015.169. URL https://doi.org/10.1109/ICCV.2015.169.

Girshick, Ross, Donahue, Jeff, Darrell, Trevor, and Malik, Jitendra. Rich feature hierarchies for accurate object detection and semantic segmentation, 2014.

Ha Q. Nguyen, Khanh Lam3, Linh T. Le et al. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. 1 2021.

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian. Deep residual learning for image recognition, 2015.

Institute, Vingroup Big Data. Vinbigdata chest x-ray abnormalities detection, 2020.

Irvin, Jeremy, Rajpurkar, Pranav, and et al., Ko. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. 33:590–597, Jul. 2019. doi: 10.1609/aaai.v33i01.3301590. URL https://ojs.aaai.org/index.php/AAAI/article/view/3834.

Johnson, Alistair EW and Pollard, Tom J et al. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1): 1–8, 2019.

Karargyris, Alexandros, Kashyap, Satyananda, Lourentzou, Ismini, Wu, Joy T., Sharma, Arjun, Tong, Matthew, Abedin, Shafiq, Beymer, David, Mukherjee, Vandana, Krupinski, Elizabeth A., and Moradi, Mehdi. Creation and validation of a chest x-ray dataset with eye-tracking and report dictation for ai development. *Scientific Data*, 2021.

Kohli, Marc, Prevedello, Luciano M, Filice, Ross W., and Geis, J. Raymond. Implementing machine learning in radiology practice and research. *American Journal of Roentgenology*, 208:754–760, 2017. URL https://www.ajronline.org/doi/full/10.2214/AJR.16.17224.

Kora Venu, Sagar and Ravula, Sridhar. Evaluation of deep convolutional generative adversarial networks for data augmentation of chest x-ray images. *Future Internet*, 13 (1), 2021. ISSN 1999-5903. URL https://www.mdpi.com/1999-5903/13/1/8.

Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E. Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Madani, A., Moradi, M., Karargyris, A., and Syeda-Mahmood, T. Semi-supervised learning with generative adversarial networks for chest x-ray classification with ability of data domain adaptation. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 1038–1042, 2018. doi: 10.1109/ISBI.2018.8363749.

Majkowska, Anna, Mittal, Sid, and Steiner, David F. et al. Chest radiograph interpretation with deep learning models: Assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. *Radiology*, 294(2):421–431, 2020. doi: 10.1148/radiol.2019191293. URL https://doi.org/10.1148/radiol.2019191293. PMID: 31793848.

Mathers, Colin D and Loncar, Dejan. Projections of global mortality and burden of disease from 2002 to 2030. *PLOS Medicine*, 2006.

McBee, Morgan P., Awan, Omer A., Colucci, Andrew T., and Ghobadi, Comeron W. Deep learning

in radiology. *Radiology Research Alliance*, 208: 754–760, 11 2018. URL https://www-clinicalkey-com.ezproxy.is.ed.ac.uk/#!/content/playContent/1-s2.0-S1076633218301041?returnurl=null&referrer=null.

Mitchell, T. M. The need for biases in learning generalizations. Technical report, Computer Science Department, Rutgers University, New Brunswick, MA, 1980.

Newell, A. and Rosenbloom, P. S. Mechanisms of skill acquisition and the law of practice. In Anderson, J. R. (ed.), *Cognitive Skills and Their Acquisition*, chapter 1, pp. 1–51. Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, 1981.

Nguyen, Ha Q., Lam, Khanh, Le, Linh T., and et al., Hieu H. Pham. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations, 2021.

Oakden-Rayner, Luke. Exploring the chestxray14 dataset: problems, Dec 2017. URL https://lukeoakdenrayner.wordpress.com/2017/12/18/the-chestxray14-dataset-problems/.

Phan, Trong Huy and Yamamoto, Kazuma. Resolving class imbalance in object detection with weighted cross entropy losses, 2020.

Reardon, Sara. Rise of robot radiologists, 2019. URL https://www.nature.com/articles/d41586-019-03847-z.

Ren, Shaoqing, He, Kaiming, Girshick, Ross, and Sun, Jian. Faster r-cnn: Towards real-time object detection with region proposal networks, 2016.

Solovyev, Roman, Wang, Weimin, and Gabruseva, Tatiana. Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing*, 107:104117, Mar 2021. ISSN 0262-8856. doi: 10.1016/j.imavis.2021.104117. URL http://dx.doi.org/10.1016/j.imavis.2021.104117.

Szegedy, Christian, Liu, Wei, Jia, Yangqing, Sermanet, Pierre, Reed, Scott, Anguelov, Dragomir, Erhan, Dumitru, Vanhoucke, Vincent, and Rabinovich, Andrew. Going deeper with convolutions, 2014.

Szegedy, Christian, Vanhoucke, Vincent, Ioffe, Sergey, Shlens, Jonathon, and Wojna, Zbigniew. Rethinking the inception architecture for computer vision, 2015.

Tan, Mingxing and Le, Quoc V. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020.

Thrall, James H., Li, Xiang, Li, Quanzheng, Cruz, Cinthia, Do, Synho, Dreyer, Keith, and Brink, James. Artificial intelligence and machine learning in radiology: Opportunities, challenges, pitfalls, and criteria for success. *Journal of the American College of Radiology*, 15:504–508, 3 2018. URL https://doi.org/10.1371/journal.pmed.0030442.

Wang, Chien-Yao, Bochkovskiy, Alexey, and Liao, Hong-Yuan Mark. Scaled-yolov4: Scaling cross stage partial network, 2021.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., and Summers, R. M. 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)*, pp. 3462–3471, 2017. doi: 10.1109/CVPR.2017.369.

Welling, Rodney D. and et al., Ezana M. Azene. White paper report of the 2010 rad-aid conference on international radiology for developing countries: Identifying sustainable strategies for imaging services in the developing world. *Journal of the American College of Radiology*, 8(8):556–562, 2011. ISSN 1546-1440. doi: https://doi.org/10.1016/j.jacr.2011.01.011. URL https://www.sciencedirect.com/science/article/pii/S1546144011000299.

Wu, Joy T., Wong, Ken C. L., and et al., Gur. Comparison of Chest Radiograph Interpretations by Artificial Intelligence Algorithm vs Radiology Residents. *JAMA Network Open*, 3(10):e2022779–e2022779, 10 2020. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2020.22779. URL https://doi.org/10.1001/jamanetworkopen.2020.22779.

Wu, Yuxin, Kirillov, Alexander, Massa, Francisco, Lo, Wan-Yen, and Girshick, Ross. Detectron2. https://github.com/facebookresearch/detectron2, 2019.