Chest X-Ray Annotation Optimisation Model Using Weighted Boxes Fusion

Kyla Joy Shitan
University of the Immaculate Conception
Davao, Philippines
kshitan_200000000995@uic.edu.ph

Julieza Jane Bella Raper
University of the Immaculate Conception
Davao, Philippines
jraper_200000000052@uic.edu.ph

Karl Vincent Bersamin
University of the Immaculate Conception
Davao, Philippines
kbersamin_200000000668@uic.edu.ph

ACM Reference Format:
Kyla Joy Shitan, Julieza Jane Bella Raper, and Karl Vincent Bersamin. 2023. Chest X-Ray Annotation Optimisation Model Using Weighted Boxes Fusion. In *Proceedings of* . ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

X-Rays are one of the hardest medical data to analyse. One such widely performed X-Ray is the CXR, or the Chest X-Ray. The interpretation of the chest radiograph can be challenging due to the superimposition of anatomical structures along the projection direction. This effect can make it very difficult to detect abnormalities in particular locations (for example, a nodule posterior to the heart in a frontal CXR), to detect small or subtle abnormalities, or to accurately distinguish between different pathological patterns [2]. It is the most widely performed X-ray and is considered to be the most difficult to interpret. All grades within a general internal medicine (GIM) team are ever reliant on the CXR, yet the ability to confidently diagnose the plethora of pathophysiology or to accurately document findings can be unreliable [9].

The volume of CXR images acquired, the complexity of their interpretation, and their value in clinical practice have long motivated researchers to build automated algorithms for CXR analysis. Indeed, this has been an area of research interest since the 1960s when the first papers describing an automated abnormality detection system on CXR images were published [2]. The potential of the automation of CXR analysis could mean better efficiency and accuracy of Chest X-Ray impressions, which in turn, increases sensitivity for subtle findings, prioritisation of time-sensitive cases, automation of tedious daily tasks, and provision of analysis in situations where radiologists are not available [2].

Efforts to automate such already exist in the world today and are optimised to be used in an actual setting, although such algorithms are trained to be able to detect a specific lung condition such as lung cancer, tuberculosis, and pneumothorax [1]. Even with the availability of these software algorithms, there still is a lingering concern about the accuracy of such programs [12].

In this study, the researchers took on the challenge of the possibility of being able to optimise the tedious process of annotating Chest X-Rays to provide medical practitioners such as radiologists and physicians with the best annotations, thus being able to conclude an accurate impression. To achieve the possibility of optimising the annotation of Chest X-Rays, the researchers will be using Weighted Boxes Fusion. The researchers believe that the results of this study, will become a stepping stone for the next generation of researchers who would be willing to undertake this field of curiosity, and as well as the medical institutions and practitioners that would be able to take benefit in the context of being able to increase efficiency and accuracy, allowing medical practitioners to be able to make sound impressions faster and better.

2 Review of Related Literature

The paper discusses the detrimental effects of noisy bounding box annotations on deep neural network performance in object detection tasks [7]. The authors provide an alternate refinement approach that enables the network to adapt the noisy datasets during training to overcome this problem. By lessening the effects of noise on the detector, particularly when the noise is more pronounced, the experimental findings show that the suggested framework greatly enhances model performance. The authors additionally note the significance of addressing the issue of noisy annotations in object detection by observing that a higher amount of noise results in a bigger training loss.

In [4], the ability to learn from multiple annotators by estimating actual labels beforehand is crucial in achieving more accurate and reliable abnormality detection. The Weighted Boxes Fusion (WBF) algorithm used in this framework is an effective way to aggregate annotations and obtain confidence values, which are then used to train a deep learning

, ,
© 2023 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

detector with a reweighted loss function. The proposed approach can help address the challenges associated with the subjectivity of human annotators and the need for large annotated datasets, which are often costly and time-consuming to create. By capturing the reliability of different annotators and incorporating confidence values into the training process, the proposed approach can significantly improve the efficiency of neural networks in anomaly detection tasks.

In [5], the researchers propose a bounding box regression loss that learns bounding box transformation and localization variance together. The KL Loss improves localization accuracy with nearly no additional computation, and the Learned Localization Variance merges neighboring bounding boxes during NMS, further improving localization performance. The researchers report significant improvements in the average precision (AP) of various architectures trained on the MS-COCO dataset.

The paper presents an ensemble algorithm that can be used with any object detection model to improve performance [3]. The algorithm has three voting strategies (affirmative, consensus, and unanimous) that are chosen based on the performance of the base models. The ensemble method is also used to define a test-time augmentation procedure for object detection that improves accuracy. The paper also discusses how the ensemble method can be used for data and model distillation, which reduces the number of annotated images needed to train a model. The methods are implemented in an open-source library and have been tested with several datasets, achieving up to a 10 percent improvement from the base models. Future work includes testing the techniques with other datasets, exploring other detection algorithms, and studying the use of the ensemble procedure as a defense against adversarial attacks.

In [10], the authors evaluated the performance of different ensemble methods to improve the performance of object detection models. The researchers proposed a new technique called WBF, which uses confidence scores of all proposed bounding boxes in an iterative algorithm that constructs the averaged boxes. The method was evaluated on two popular benchmark datasets for object detection, and it outperformed the other methods for combining predictions and achieved top results in the Open Images Detection Challenge and the COCO Detection Challenge. The authors' approach is straightforward and can be easily applied to other object detection models and datasets.

This study discusses the use of the WBF algorithm for ensembling in lung disease detection [13]. The study showed that the proposed method can significantly improve the detection capacity of neural networks in lung disease detection tasks. The authors' experimental results identified the superiority of the WBF algorithm and the proposed method for the efficacy of the disease detection task.

Additionally, a study on the detection of Pneumonia in chest X-ray images, as a way to assist doctors, explores the Non-Maximum Suppression method [14]. The researchers proposed a deep learning method based on RetinaNet for pneumonia detection. They introduced Res2Net into RetinaNet to get a multi-scale feature of pneumonia, followed by the proposition of novel predicted boxes fusion algorithm, named Fuzzy-Non Maximum Suppression (FNMS). FNMS works by fusing the overlapping detection boxes, providing a more robust prediction area. The model's performance, according to the researchers, outperforms existing methods because of the integration of two models with different backbones. The researchers report the results in the single model case and the model ensemble case, where the predicted boxes that were fused by the FNMS algorithm outperformed other fusion methods. [8] presents a novel approach to using NMS by adapting Affinity Propagation Clustering (APC) to remove false positives. The researchers use the windows proposed by the object detector as data points for APC and propose the use of Latent Structured SVM (LSSVM) to learn the weights of APC.

Similarly, [6] addresses misdetection and false object detection in a Single Threshold-NMS algorithm by proposing a GDT-NMS algorithm that uses Generalized Intersection over Union (GIoU) to compute the similarity between objects. The dual-threshold NMS algorithm balances the relationship between misdetection and false object detection, resulting in better detection results compared to the NMS algorithm.

In [11], the researchers propose a line-segment-based NMS method that utilizes the distribution of line segments and multiple differentiated metrics to integrate geometric information onto existing models. The method overcomes the limitations of area-based metrics by using the intrinsic information of contour edges, facilitating the localization of bounding boxes and achieving more precise detections.

Overall, these articles suggest that enhancing the NMS and WBF technique can significantly improve the accuracy and efficiency of object detection algorithms. The proposed methods provide various ways to integrate additional information, such as geometric information and similarity measures, to overcome the limitations of existing methods. These approaches may help address the challenges of misdetection, false object detection, and duplicated detections commonly encountered in object detection models.

3 Methodology

The methodology used in this study is Experimental Research Methodology. The dataset used in this is the VinBigData Chest X-ray Abnormalities.

Since there are repeating annotations due to different radiologists annotating the images, there will be overlaps. To solve this issue, ensemble boxes are used to combine the bounding boxes generated by different radiologists. The ensemble boxes are generated by taking the union of all bounding boxes of the same object detected in the same image.

The aim of this study is to find the most efficient method for object detection and localization in chest X-ray images. To achieve this, different methods are tested and compared. The methods tested are Non-maximum Suppression (NMS), Soft-NMS, Non-maximum Weighted (NMW), and Weighted Bboxes Fusion (WBF).

Overall, the aim of this study is to identify the most efficient method for object detection and localization in chest X-ray images, which can potentially improve the accuracy of clinical diagnoses and treatment.

3.1 Data Collection

The VinBigData Chest X-ray Abnormalities dataset used in this study consists of 15000 postero-anterior (PA) chest X-ray (CXR) scans in DICOM format. The dataset is divided into two folders: training and testing, with 15000 and 3000 images, respectively. Each image in the dataset was labeled by a panel of 17 experienced radiologists for the presence of 14 critical radiographic findings. These are the following abnormalities along with their definition:

0 - Aortic enlargement: An abnormal condition in which the aorta, the body's biggest blood vessel that delivers blood from the heart to the rest of the body, enlarges.

1 - Atelectasis: An abnormality that refers to the partial or complete collapse of a lung or a lobe of the lung.

2 - Calcification: The accumulation of calcium deposits in the body tissues.

3 - Cardiomegaly: Enlargement of the heart, which can be caused by a variety of factors, including high blood pressure and heart disease.

4 - Consolidation: Replacement of air in the lungs with fluid, leading to increased opacity in the lung tissue.

5 - ILD (Interstitial Lung Disease): This refers to a group of lung diseases that affect the tissue and space around the air sacs of the lungs.

6 - Infiltration: The presence of abnormal substances or cells within the lung tissue, such as pus, blood, or cancer cells.

7 - Lung Opacity: The presence of an abnormal shadow or area of increased density on a chest X-ray, which can indicate a variety of conditions, including fluid in the lungs or a tumor.

8 - Nodule/Mass: The presence of a round or oval-shaped lesion in the lung tissue, which can be benign or malignant.

9 - Other lesion: This refers to any abnormality that does not fit into one of the other categories.

10 - Pleural effusion: Buildup of excess fluid between the layers of tissue that line the lungs and chest cavity, which can cause difficulty breathing.

11 - Pleural thickening: This condition, which can be brought on by a number of things like inflammation and scarring, refers to the thickening of the tissue that lines the lungs and chest cavity.

12 - Pneumothorax: This condition refers to the presence of air between the chest wall and the lung, which might result in the lung collapsing.

13 - Pulmonary fibrosis: This is a disorder that causes the lung tissue to thicken and scar, makes breathing difficult.

The dataset also includes a train.csv file that contains metadata for each object that was identified in the labeled photos, including the class and bounding box details. The file has eight columns which are the image_id, class_name, class_id, rad_id, x_min, y_min, x_max, and y_max. The column image_id serves as a unique identifier for each image. The class_name gives the name of the detected object. This includes the 14 abnormalities and the class “no findings” if no object was detected. The class_id gives the numerical ID of the detected object. The radiologist's ID is represented by the rad_id. Then there are the remaining four columns: x_min, y_min, x_max, and y_max which represent the minimum and maximum coordinates of the bounding box surrounds the object. These objects are listed in different rows in the train.csv file.

3.2 Data Cleaning and Transformation

The Data cleaning and transformation process involved several steps. First, we converted the DICOM file format to JPEG to reduce the file size from 206 GB to 3.4 GB for easier processing. Next, as the class "No finding" did not have minimum and maximum bounding box coordinates, we next eliminated it from the dataset. And due to the image being downsized, the bounding box's coordinates also required to be changed. To do this, we reduced the values forx_min, y_min, x_max, and y_max so that they matched the reduced sizes of the photos. This procedure helped in making sure the data could be used for the subsequent analysis and development.

3.3 Data Analysis

Based on the dataset, we have identified that 4394 images out of the total of 18000 images contain abnormalities, with a total of 36096 annotations. When we analyzed the unique annotation counts per class, we found that aortic enlargement has the highest count, while pneumothorax has the smallest count.

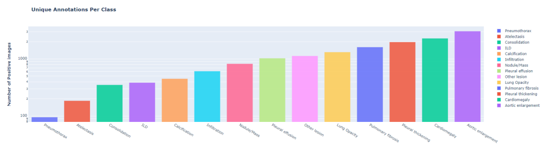

Figure 1. Annotations Per Class

Further analysis of the dataset shows that the number of annotations per radiologist varies significantly. Radiologist r9 has the highest number of annotations, while radiologist r17 has the lowest number of annotations.

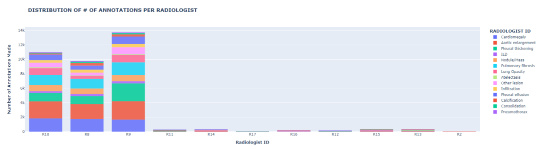

Figure 2. Annotations Per Radiologist

To visually represent the dataset, we have included an example chest X-ray image with bounding boxes based on the CSV file with their respective abnormalities. The image shows how multiple abnormalities can exist within one image, and the location of each abnormality is clearly marked with a bounding box and labeled with their respective class name.

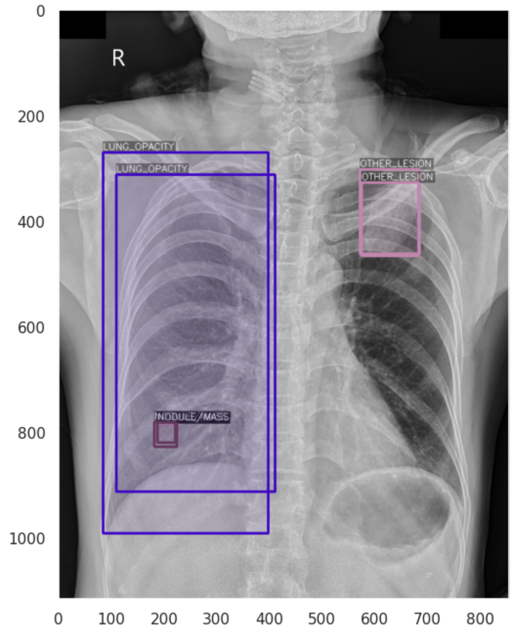

Figure 3. Example Annotation

3.4 Development

This code implements four different algorithms for object detection: Non-maximum Suppression (NMS), Soft-NMS, Weighted Boxes Fusion, and Non-maximum Weighted (NMW). Non-maximum Suppression (NMS) is a post-processing technique used in object detection to eliminate overlapping bounding boxes. Soft-NMS is an extension of NMS, where instead of removing overlapping bounding boxes, the scores of the boxes are decayed based on their overlap with other boxes. Non-maximum Weighted (NMW) is another extension of NMS, where the boxes are weighted based on their overlap and then merged. Weighted Bboxes Fusion (WBF) is a fusion method that combines multiple detections from different models or algorithms.

The difference between these algorithms is the way they filter out overlapping bounding boxes. The code loads an image and its annotations and then performs object detection on the image. For each image, it performs Non-maximum Weighted (NMW) to filter out overlapping bounding boxes. It then converts the ground truth annotations and predicted bounding boxes to tensors, calculates the Intersection over Union (IoU) between them, and stores true positives, false positives, and ground truths to calculate the F1-score. Finally, it visualizes the original bounding boxes, predicted bounding boxes, and overlapping bounding boxes on the image.

3.5 Evaluation

To evaluate the methodology, we have calculated various performance metrics such as precision, recall, average precision, F1 score, average IoU, and average recall. These metrics are commonly used in object detection tasks to assess the accuracy and effectiveness of the algorithm.

The precision metric measures the proportion of true positive predictions out of all positive predictions, while the recall metric measures the proportion of true positive predictions out of all true positives in the ground truth. The average precision metric is calculated by averaging the precision values at different recall levels.

We have also calculated the F1 score, which is the harmonic mean of precision and recall. This metric is used when you want to balance the importance of precision and recall in the evaluation of the algorithm.

Furthermore, we have calculated the average IoU, which is the mean of the maximum intersection over union (IoU) values between the predicted bounding boxes and the ground truth bounding boxes. This metric is commonly used to evaluate the accuracy of object localization.

Lastly, we have also calculated the average recall, which is the mean of the recall values at different IoU thresholds. This metric is used to evaluate the ability of the algorithm to detect objects at different levels of IoU overlap.

References

[1] Google AI Blog. 2021. Detecting Abnormal Chest X-Rays Using Deep Learning. *Google AI Blog* (September 2021). https://ai.googleblog.com/2021/09/detecting-abnormal-chest-x-rays-using.html

[2] Erdi Çallı, Ecem Sogancioglu, Bram van Ginneken, Kicky G van Leeuwen, and Keelin Murphy. 2021. Deep learning for chest X-ray analysis: A survey. *Medical Image Analysis* 72 (2021), 102125.

[3] Ángela Casado-García and Jónathan Heras. 2020. Ensemble Methods for Object Detection. In *European Conference on Artificial Intelligence*.

[4] Khiem H. Le, Tuan V. Tran, Hieu H. Pham, Hieu T. Nguyen, Tung T. Le, and Ha Q. Nguyen. 2023. Learning From Multiple Expert Annotators for Enhancing Anomaly Detection in Medical Image Analysis. *IEEE Access* 11 (2023), 14105–14114. https://doi.org/10.1109/ACCESS.2023.3243845

[5] Yihui He, Chenchen Zhu, Jianren Wang, Marios Savvides, and Xiangyu Zhang. 2019. Bounding Box Regression With Uncertainty for Accurate Object Detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2883–2892. https://doi.org/10.1109/CVPR.

2019.00300

[6] Zhiqiang Hou, Xiaoyi Liu, and Lilin Chen. 2020. Object Detection Algorithm for Improving Non-Maximum Suppression Using GIoU. *IOP Conference Series: Materials Science and Engineering* 790 (04 2020), 012062. https://doi.org/10.1088/1757-899X/790/1/012062

[7] Jiafeng Mao, Qing Yu, Yoko Yamakata, and Kiyoharu Aizawa. 2021. Noisy Annotation Refinement for Object Detection.

[8] Rasmus Rothe, Matthieu Guillaumin, and Luc Van Gool. 2015. Non-Maximum Suppression for Object Detection by Passing Messages between Windows. *LNCS* 9003. https://doi.org/10.1007/978-3-319-16865-4_19

[9] I Satia, S Bashagha, A Bibi, R Ahmed, S Mellor, and F Zaman. 2013. Assessing the accuracy and certainty in interpreting chest X-rays in the medical division. *Clinical medicine* 13, 4 (2013), 349.

[10] Roman Solovyev, Weimin Wang, and Tatiana Gabruseva. 2021. Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing* 107 (2021), 104117. https://doi.org/10.1016/j.imavis.2021.104117

[11] Xue-Song Tang, Xianlin Xie, Kuangrong Hao, Dawei Li, and Mingbo Zhao. 2022. A line-segment-based non-maximum suppression method for accurate object detection. *Knowledge-Based Systems* 251 (05 2022), 108885. https://doi.org/10.1016/j.knosys.2022.108885

[12] Gamuchirai Tavaziva, Miriam Harris, Syed K Abidi, Coralie Geric, Marianne Breuninger, Keertan Dheda, Aliasgar Esmail, Monde Muyoyeta, Klaus Reither, Arman Majidulla, et al. 2022. Chest X-ray analysis with deep learning-based software as a triage test for pulmonary tuberculosis: an individual patient data meta-analysis of diagnostic accuracy. *Clinical Infectious Diseases* 74, 8 (2022), 1390–1400.

[13] Huy Tran, Long TonThat, and Kien Trang. 2022. Weighted Box Fusion Ensembling for Lung Disease Detection. In *8th International Conference on the Development of Biomedical Engineering in Vietnam: Proceedings of BME 8, 2020, Vietnam: Healthcare Technology for Smart City in Low- and Middle-Income Countries*. Springer, 743–750.

[14] Hongli Wu, Mingzhu Ping, Huijuan Lu, and Wenjie Zhu. 2021. A Deep Learning Method for Pneumonia Detection Based on Fuzzy Non-Maximum Suppression. In *2021 2nd International Conference on Artificial Intelligence and Computer Engineering (ICAICE)*. 89–94. https://doi.org/10.1109/ICAICE54393.2021.00026