Detection of lesions in chest radiographs

Weronika Hryniewska
weronika.hryniewska.dokt@pw.edu.pl
Faculty of Mathematics and Information Science,
Warsaw University of Technology

Abstract

The work based on a Kaggle competition is related to the detection of abnormalities in X-ray lung images. Complex models are able to learn important features as well as artifacts, and it is difficult to remove this bias during or after the training. This single dataset scenario can be adopted to similar problems.

1. Introduction

Radiography (X-ray) is an imaging technique that uses a small dose of ionizing radiation to create images of the internal structures of a body. Due to the relatively low price of the device and the existence of portable devices, X-ray imaging is a widely used technique. However, it is particularly difficult to assess the severity of the pathology, and, thus, only experts in radiology should interpret chest images.

Recent applications of machine learning (ML) have gained popularity in the medical domain [14, 2]. The performance achieved by neural networks is becoming similar to that reached by medical experts [15].

Considering the need for a highly precise and fast diagnosis process, on the Kaggle platform was announced a competition about automatically localizing and classifying thoracic abnormalities from chest radiographs [6]. On December 30, 2020, the database with 18,000 posterior-anterior (PA) X-ray scans in DICOM format became available on: [12]. More than 1,300 teams were participating in the competition trying to train the best model. The total prize money in this challenge was 50,000 dollars. The crucial value of the dataset was in the annotations. They were created by radiologists and show the location of anomalies in chests.

In this paper, we will create a model for localization and classification task called detection task.

2. The training set

The training set contains 15,000 lung images in DICOM format with annotations. Each image was annotated by three radiologists. Due to a DICOM format, the images have high quality and the information about a patient (such as age or sex) or about the image (such as the number of allocated bits) is included.

There are fourteen labels for lesions and one additional label for images of healthy lungs: Aortic enlargement, Atelectasis, Calcification, Cardiomegaly, Consolidation, ILD, Infiltration, Lung Opacity, Nodule/Mass, Pleural effusion, Pleural thickening, Pneumothorax, Pulmonary fibrosis, Other lesions, No finding. Examples of bounding box visualizations are presented in Figure 1.

The test set has 3,000 DICOM files with no annotations.

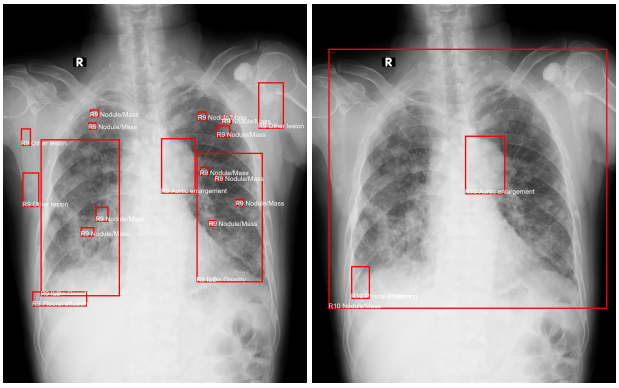

Figure 1. Bounding boxes visualizations

2.1. Different procedure of preparing train and test sets

The train and test sets were prepared differently. In both, the annotations were made independently by three radiologists for each image. According to [6], in the test set, there was an additional processing step. The labels were additionally verified and a consensus between two radiologists was reached.

The problem is that there are considerable differences between radiologists. One approach is to select only critical findings and discard other annotations as unnecessary, which is acceptable for radiologists, but very challenging for nowadays ML model architectures. Typically, there is an assumption that a ML model should be trained on data similar to the target, and in order to deal with noise, more data is required.

The role of two expert radiologists is unclear. It seems that those two only corrected annotations made by others, while their role should be much bigger.

2.2. Data quality

Two monochromatic color spaces Another valid concern is Photometric Interpretation, which specifies the intended interpretation of the image pixel data. Some images are of type *monochrome1* (17%) and some of *monochrome2*, some examples are shown in Figure 2.. The difference is that in the first case the lowest value of a pixel is interpreted as white and in the second case as black. This may produce some inefficient models when not taken into consideration.

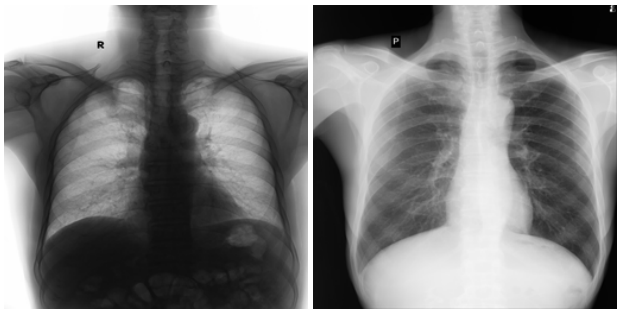

Figure 2. Example of monochrome1 and monochrome2

Parts of clothes present in the X-rays. Undesirable artifacts are present in many images, which in some cases can reduce the diagnostic value of the image, and when used for machine learning, introduce additional noise. Some examples are shown in Figure 3. These artifacts can be easily avoided during image acquisition, by asking the patient to remove all parts of the clothes that may influence X-ray imaging, for example, chains, bras, clothes with buttons, and zippers. If artifacts cannot be prevented, they can be removed during image preprocessing, before the image is shown to the model.

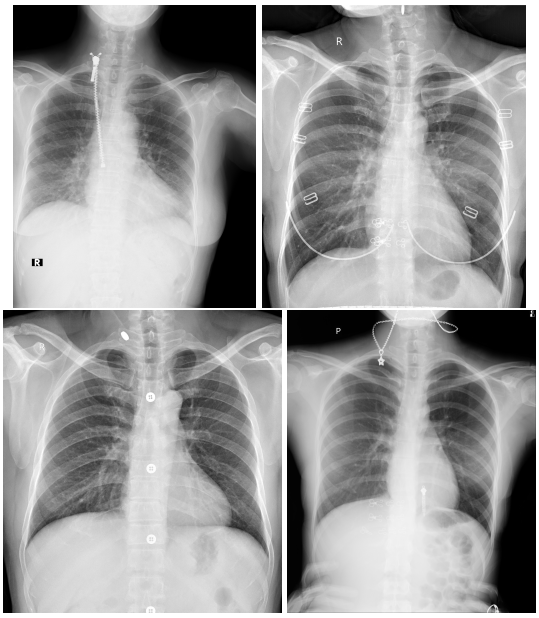

Figure 3. Example of clothes artifacts. From the left, there are: a zipper, a bone in a bra, buttons, and a chain.

Letters and annotations present in the X-rays Letters and/or annotations present in some lung images should be removed during preprocessing to prevent a neural network from learning those patterns. The model should learn how to differentiate labels by focusing on image features, not on descriptions in the images.

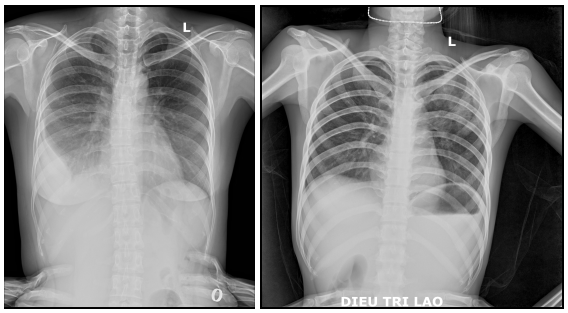

Figure 4. Example of other artifacts. From the left, there are: letters, and annotations.

3. Model

3.1. Image preprocessing

Image preprocessing was performed following the guidelines from: [10], and [7]. In addition, we tested also two very popular methods: histogram equalization and image normalization.

The first paper described the positive influence of proper medical data preparation for the automatic detection of lung nodules. For our purposes, we added Gabor filter as an appropriate lung image preprocessing method. Alpha-

trimmed mean filter appeared to drastically degrade the image quality, so we decided to remove it.

The second one reviewed the AutoML techniques with a surprising discovery that adding Gaussian noise to the image improves a neural network's performance and robustness.

However, ablation study ended with conclusion that the best results are obtained by the neural network without any additional image preprocessing. We assume that it is caused by greatly reduced input image dimensions, and for this reason any additional image preprocessing may cover valuable information about the lesions location.

3.2. Data augmentation

In our research, the following augmentation techniques were used: scaling up to 15%, rotating up to 10°, random brightness and contrast, adding more Gaussian noise with mean value 0 and variance range for noise between 10 and 50, and horizontal flip. The last one is rather not recommended due to the medical domain. However, the possibility of imbalance in right and left lung annotation forced us to used it. Every augmentation was performed with probability 50%.

Unfortunately, despite their high efficiency, it appears that Grid Distortion and Elastic Transform are not available for neural networks that contain bounding boxes.

3.3. Model architecture and training

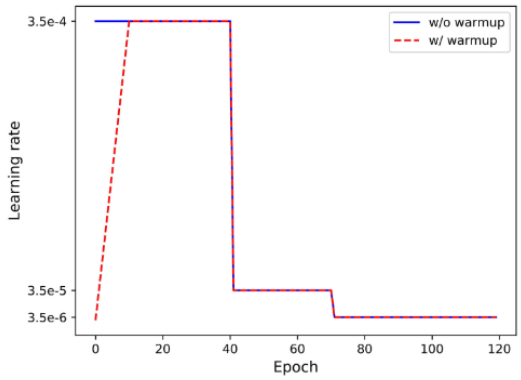

Figure 6. Comparison between two learning rate schedules. Warmup MultiStep Learning Rate scheduler increased the learning rate linearly during at the beginning of training. [5]

We created Faster R-CNN on Feature Pyramid Network (FPN) with ResNet-50 backbone using Detectron2 library [13]. The input image size is 256x256. Faster R-CNN detector with FPN backbone is a multi-scale detector that detects tiny as well as large objects in the images with high accuracy. The proposed architecture is presented in the Figure 5.

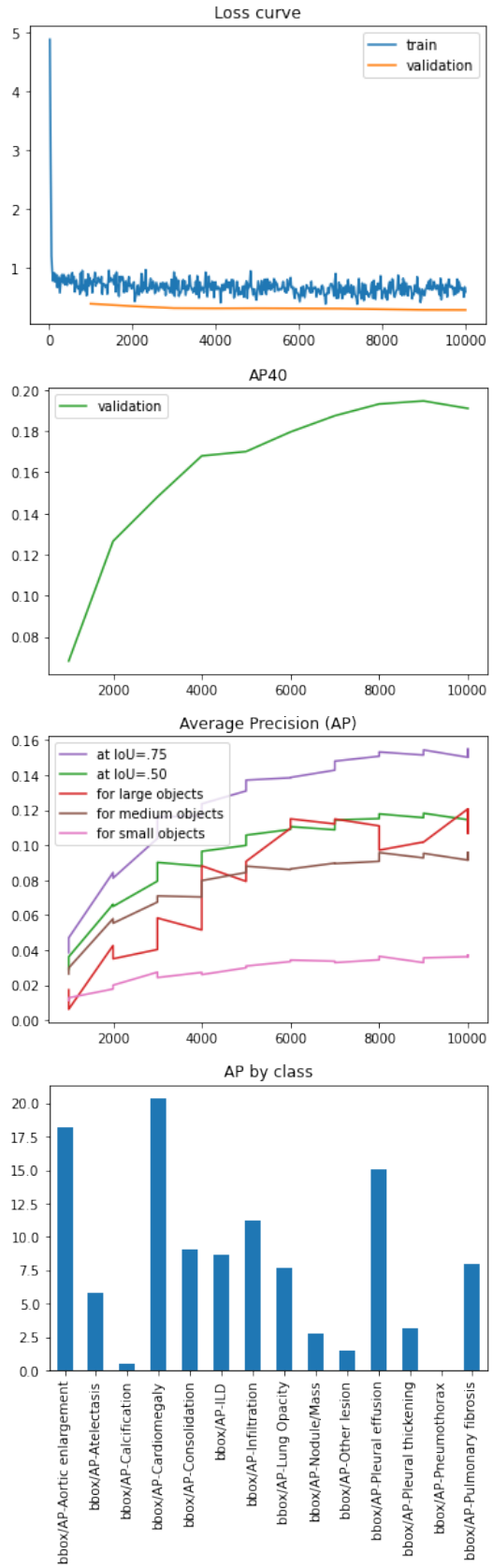

Figure 7. Loss and average precision versus epochs graphs

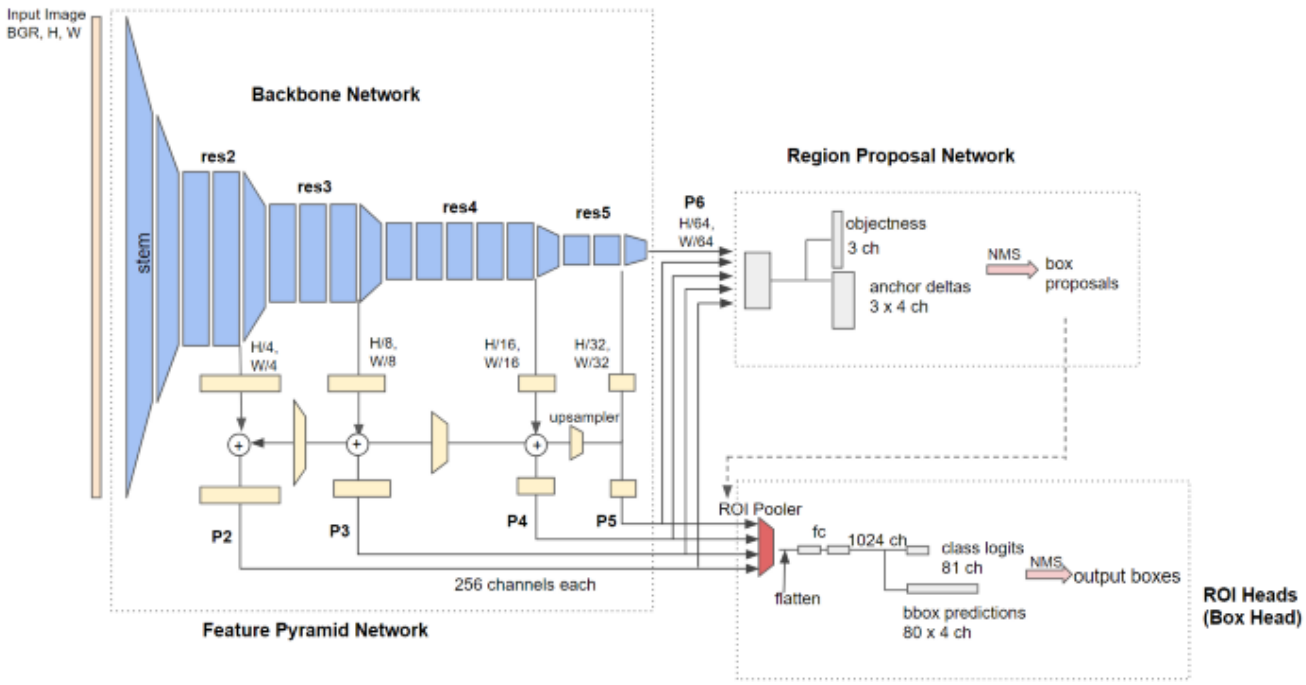

Figure 5. Detailed architecture of Faster-RCNN-FPN [3]

The loss curve of the designed solution is presented in Figure 7. It is visible that the neural network achieves satisfactory results quite quickly and that the training is stable.

To adjust the learning rate at the beginning of a training Warmup MultiStep Learning Rate scheduler was chosen, Figure 6.

3.4. Metrics

The metric defined by competition organizers involves a detection of regions with potential lesions with mean Average Precision (mAP) at Intersection over Union (IoU) > 0.4 [1]. Our neural network achieves 19.5%.

Mean Average Precision is a popular metric to measure the performance in object detection. Intersection over Union at 0.4 means that the predicted box will be considered as true positive if the quotient of the area of overlap and the area of union will be greater than 40%.

In Figure 7 in the third chart, there is a differentiation of the average precision for the object of different sizes. As small objects are considered objects below 32^2 pixels. The area of medium objects are between 32^2 and 96^2 pixels. Objects bigger than 96^2 pixels are big. The most difficult to recognize were small objects. For this reason, there is a great possibility that if the image dimensions are bigger, the small object recognition will be easier.

The fourth chart shows that the more typical location of a lesion, the better recognition result. Cardiomegaly indicates that the heart of the person examined is enlarged in relation to its natural size, whereas aortic enlargement points out to a patient's aorta. This fact might be not surprising unless we take into account two different X-ray projections. The neural network had to learn the relation between a body size to the heart size in both projections: Anterior-Posterior (AP) and Posterior-Anterior (PA).

4. Visualizations

Explanations of decisions taken by deep neural networks started to be crucial in the medical field. COVID-19 pandemic forced researchers to pay even more attention to provide interpretable output from the developed models [4].

However, surprisingly, even in systematic reviews of Explainable Artificial Intelligence [11], there is a lack of description of methods designed for detection problems. The researchers focus mainly on classification and regression problems.

Due to the fact that the detection is a classification task with additional localization information, there was an attempt to use standard methods, such as Gradient-weighted Class Activation Mapping (Grad-CAM) [9] and Local Interpretable Model-Agnostic Explanations (LIME) [8].

Unfortunately, the currently available implementations do not allow to use them for a detection task. A lack of explainable methods' implementation for detection tasks leaves the created model "black-box". Without proper explainable methods, it is difficult to validate the model correctly.

5. Conclusions and recommendations

The quality of a model is inherently bound to the quality of the data on which it is trained. Developing of a reliable model should begin with proper data acquisition (without artifacts and descriptions) and annotation (consistent annotation rules). At the model development stage, we cannot make the model fulfill all responsible AI rules if the data and their annotations are of insufficient quality, and there is a lack of dedicated explanation methods.

The main problem faced in this project was a lack of proper explanation methods. As a recommendation, we would like to suggest to focus more on preparing such easily used, dedicated methods.

6. Acknowledgments

We would like to thank the authors of publicly available notebooks who shared their work to help others in designing their solutions.

References

[1] Mark Everingham, Luc Van Gool, Christopher K.I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, jun 2010.

[2] Yuyu Guo, Lei Bi, Euijoon Ahn, Dagan Feng, Qian Wang, and Jinman Kim. A spatiotemporal volumetric interpolation network for 4d dynamic medical image. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[3] Hiroto Honda. Digging into Detectron 2. part 1. `https://medium.com/@hirotoschwert/digging-into-detectron-2-47b2e794fabd`.

[4] Weronika Hryniewska, Przemysław Bombiński, Patryk Szatkowski, Paulina Tomaszewska, Artur Przelaskowski, and Przemysław Biecek. Checklist for responsible deep learning modeling of medical images based on covid-19 detection studies. *Pattern Recognition*, page 108035, 2021.

[5] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of Tricks and A Strong Baseline for Deep Person Re-identification. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2019-June:1487–1495, mar 2019.

[6] Ha Q. Nguyen, Khanh Lam, Linh T. Le, Hieu H. Pham, Dat Q. Tran, Dung B. Nguyen, Dung D. Le, Chi M. Pham, Hang T. T. Tong, Diep H. Dinh, Cuong D. Do, Luu T. Doan, Cuong N. Nguyen, Binh T. Nguyen, Que V. Nguyen, Au D. Hoang, Hien N. Phan, Anh T. Nguyen, Phuong H. Ho, Dat T. Ngo, Nghia T. Nguyen, Nhan T. Nguyen, Minh Dao, and Van Vu. VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations. dec 2020.

[7] Esteban Real, Chen Liang, David R. So, and Quoc V. Le. AutoML-Zero: Evolving Machine Learning Algorithms From Scratch. *37th International Conference on Machine Learning, ICML 2020*, PartF168147-11:7963–7975, mar 2020.

[8] Marco Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101. Association for Computational Linguistics (ACL), jul 2016.

[9] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2017-October, pages 618–626. Institute of Electrical and Electronics Engineers Inc., dec 2017.

[10] Faridoddin Shariaty and Mojtaba Mousavi. Application of cad systems for the automatic detection of lung nodules. *Informatics in Medicine Unlocked*, 15:100173, 2019.

[11] Giulia Vilone and Luca Longo. Explainable Artificial Intelligence: a Systematic Review. may 2020.

[12] Vingroup Big Data Institute. VinBigData Chest X-ray Abnormalities Detection — Kaggle. `www.kaggle.com/c/vinbigdata-chest-xray-abnormalities-detection/`, urldate = 2021-03-16, year = 2020.

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

[14] Qihang Yu, Dong Yang, Holger Roth, Yutong Bai, Yixiao Zhang, Alan L. Yuille, and Daguang Xu. C2fnas: Coarse-to-fine neural architecture search for 3d medical image segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[15] Yi Zhou, Xiaodong He, Lei Huang, Li Liu, Fan Zhu, Shanshan Cui, and Ling Shao. Collaborative learning of semi-supervised segmentation and classification for medical images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.