Project Proposal
Hygieia, Project A, Master AI Informatics, 2020-2021

1. Background

Among the medical imaging examinations, chest X-rays are the most common. Each year, there are 3.6 billion medical procedures involving ionizing radiation ([1]Draelos, 2019), of which over 2 billion are chest X-rays ([2]Phillips et al., 2020). Pneumonia is one among the many diseases that chest X-ray can help to detect. Every year in the US, pneumonia causes hospitalization of more than 1 million adults, and kills around 50 000 ([3]Rajpurkar et al., 2017). Detecting pneumonia is a difficult task which requires expertise of radiologists. However, there is always a shortage of radiologists to interpret the X-ray results ([2]Phillips et al., 2020). Moreover, fatigue due to heavy workload may also deteriorate the diagnostic accuracy of radiologists ([4]Rajpurkar et al., 2017). Automatic interpretation of chest X-ray would bring numerous substantial benefits, but it is also a challenging task. Lung opacity may be vague and very similar to minor benign anomalies. Variation in radiation settings may also make an image look more hazy in general. In the CheXNet paper, the authors ([3]Rajpurkar et al., 2017) reported results for the Chest X-ray 14 dataset ([10]Wang et al., 2017). Their AUROC metric ranges between 0.7 and 0.8 for the 3 classes associated with pneumonia (Pneumonia, Consolidation, Infiltration).

2. Material & method:

2.1 Data

The original dataset is RSNA Pneumonia Detection Challenge 2018 ([5]RSNA, 2018). This dataset contains publicly available chest X-ray from the NIH (National Institutes of Health). Board-certified radiologists located the opacity region with bounding boxes.

There are frontal view chest X-ray of 26 684 unique patients in the original collection. Each image belongs to one of 3 categories: Normal, Lung opacity, No lung opacity / Not normal. The Normal class contains healthy lungs. Lung opacity contains potential pneumonia patients (signs of Infiltration or Consolidation). No Lung opacity / Not normal corresponds to radiographs without opacity related to pneumonia. This class still contains other types of opacity (e.g. nodule, mass) and other kinds of anomalies. To prepare a preprocessed data version, we extract the activation from the last convolution layer of a convolution neural network.

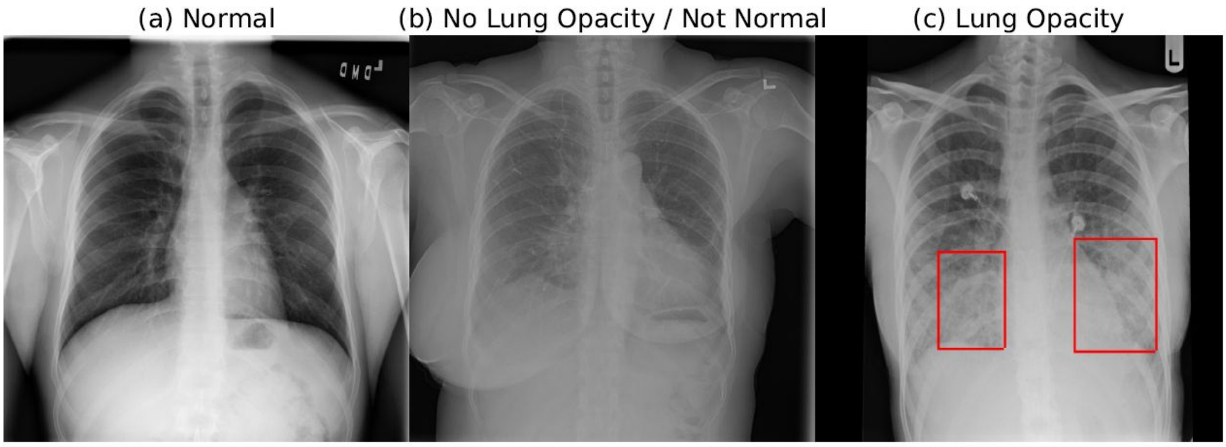


Image source ([6]Gabruseva)

2.2 Classification

Starting from the RSNA Pneumonia dataset, we propose a binary classification task of normal vs pneumonia. We select the normal images from the original collection as the negative class, and the lung opacity examples as the positive class. We decide to exclude the No lung opacity / Not normal category, since this class has a lot of variations. Another reason is to limit the size of the input data.

The training set contains around 10 000 images in total. Each of the 2 evaluation phases has 2000 images. Each unique patient has only one image, and there is no overlap between the 3 subsets. The ratio of normal to lung opacity in each subset is around 3:2. We resize all selected images from 1024 x 1024 px to 512 x 512 px. Section preliminary result shows a comparison of model performance with different image sizes.

2.3 Evaluation

In medical settings, an incorrectly predicted healthy case (= false negative) is more harmful than an incorrectly predicted ill case (= false positive). For this reason, we use f-beta score with beta = 2 (beta = weight of recall / precision). Thus, Recall is considered 2 times more important than Precision.

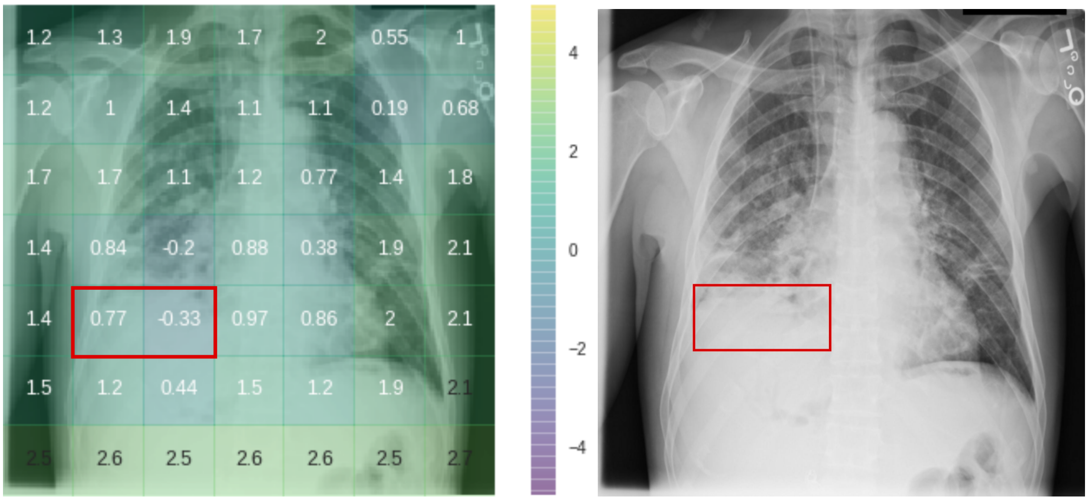
$$F_\beta = (1 + \beta^2) \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{(\beta^2 \cdot \mathrm{precision}) + \mathrm{recall}}.$$

Image from ([11]fbeta)

However, accuracy alone is not enough as assessment. John Zech ([7]Zech) demonstrated the remarkable ability of deep convolutional neural networks to detect confounding factors and take advantage of them. Hence, we suggest calculating an explanation score as a measure of trustworthiness.

We generate a class activation heatmap for a set of explanation images, using the GradCAM method ([8]Selvaraju et al., 2016). Inspired by section 8.1 of the GradCAM paper, we leverage the provided bounding box to assess the quality of the GradCAM heatmap. We define a measure of heatmap Trustworthiness as:

Trustworthiness = Activation inside box / Total activation
E.g. if this ratio = 0.8, this means activation inside the bounding box accounts for 80% of total activation. A trustworthy heatmap should have high activation inside the box, and weak or almost 0 activation elsewhere. Hence, this ration will be close to 1. Then, we simply average the trustworthiness score over all heatmaps :
Explanation score = Average(Trustworthiness of all heatmaps)



Modified from [7]Zech

3. Preliminary results

The table below shows the results with different image sizes and models.

Model	Image size	Accuracy
Mobilenetv2	128	0.87
Mobilenetv2	192	0.89
Mobilenetv2	256	0.9
Mnasnet	128	0.64

The trustworthiness score on a small sample is 0.06. A score of 1 means total similarity between the annotated bounding box and the heatmap. Whereas a score of 0 means total difference.

Hygieia group members:
Phan Anh Vu, Alejandro de la Cruz López, Paavo Camps, Xinwen Xu, Nour Jemli

Github URL:
https://github.com/phanav/upsaclay-hygieia

References

[1]Draelos. (2019). *Radiology: Normal Chest X-Rays*. Rachel Draelos. https://glassboxmedicine.com/2019/02/10/radiology-normal-chest-x-rays/

[2]Phillips et al. (2020). *CheXphoto: 10,000+ Photos and Transformations of Chest X-rays for Benchmarking Deep Learning Robustness*. Nick A. Phillips, Pranav Rajpurkar, Mark Sabini, Rayan Krishnan, Sharon Zhou, Anuj Pareek, Nguyet Minh Phu, Chris Wang, Mudit Jain, Nguyen Duong Du, Steven QH Truong, Andrew Y. Ng, Matthew P. Lungren. https://arxiv.org/abs/2007.06199

[3]Rajpurkar et al. (2017). CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. *Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, Andrew Y. Ng*. https://arxiv.org/abs/1711.05225

[4]Rajpurkar et al. (2017). *MURA: Large Dataset for Abnormality Detection in Musculoskeletal Radiographs*. Pranav Rajpurkar, Jeremy Irvin, Aarti Bagul, Daisy Ding, Tony Duan, Hershel Mehta, Brandon Yang, Kaylie Zhu, Dillon Laird, Robyn L. Ball, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, Andrew Y. Ng. https://arxiv.org/abs/1712.06957

[5]RSNA. (2018). *Augmenting the National Institutes of Health Chest Radiograph Dataset with Expert Annotations of Possible Pneumonia* (https://doi.org/10.1148/ryai.2019180041 ed.). George Shih , Carol C. Wu, Safwan S. Halabi, Marc D. Kohli, Luciano M. Prevedello, Tessa S. Cook, Arjun Sharma, Judith K. Amorosa, Veronica Arteaga, Maya Galperin-Aizenberg, Ritu R. Gill, Myrna C.B. Godoy, Stephen Hobbs, Jean

Jeudy, Archana Laroia, Palmi N. https://www.rsna.org/education/ai-resources-and-training/ai-image-challenge/RSNA-Pneumonia-Detection-Challenge-2018

[6]Gabruseva. (n.d.). *https://github.com/tatigabru/kaggle-rsna*.

[7]Zech. (n.d.). *https://jrzech.medium.com/what-are-radiological-deep-learning-models-actually-learning-f97a546c5b98*.

[8]Selvaraju et al. (2016). *Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization*. Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra. https://arxiv.org/abs/1610.02391

[9]kaggle. (n.d.). *https://www.kaggle.com/nikhilikhar/classification-with-83-accuracy*.

[10]Wang et al. (2017). *ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases*. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. https://openaccess.thecvf.com/content_cvpr_2017/papers/Wang_ChestX-ray8_Hospital-Scale_Chest_CVPR_2017_paper.pdf

[11]fbeta. (n.d.). https://en.wikipedia.org/wiki/F-score