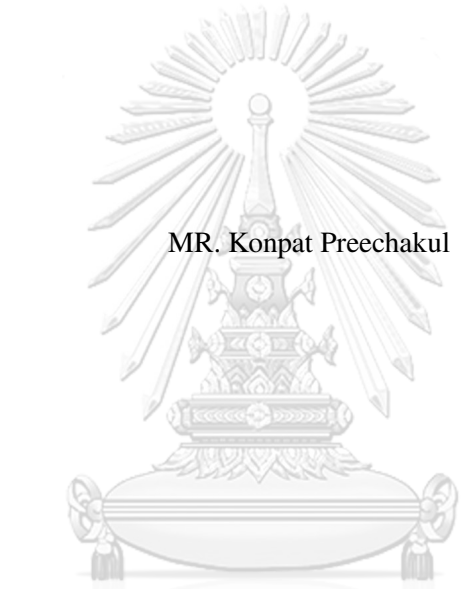
REDESIGNING WEAKLY SUPERVISED LOCALIZATION ARCHITECTURES FOR MEDICAL IMAGES

MR. Konpat Preechakul



A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Engineering Program in Computer Engineering
Department of Computer Engineering
Faculty of Engineering
Chulalongkorn University
Academic Year 2020

การออกแบบสถาปัตยกรรมหาตำแหน่งแบบสอนอย่างอ่อนสำหรับภาพทางการแพทย์

นายกรพัฒน์ ปรีชากุล

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตร์มหาบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมศาสตร์คอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2563

Thesis Title	REDESIGNING WEAKLY SUPERVISED LOCALIZATION ARCHITECTURES FOR MEDICAL IMAGES
By	MR. Konpat Preechakul
Field of Study	Computer Engineering
Thesis Advisor	Prof. Boonserm Kijsirikul, D.Eng.
Thesis Co-advisor	Ekapol Chuangsuwanich, Ph.D., Sira Sriswasdi, Ph.D.

Accepted by the Faculty of Engineering, Chulalongkorn University in Partial Fulfillment of the Requirements for the Master's Degree

Dean of the Faculty of Engineering

. . . . . .

(Prof. Supot Teachavorasinskun, D.Eng.)

THESIS COMMITTEE

. . . . . Thesis Advisor

(Prof. Boonserm Kijsirikul, D.Eng.)

. . . . . First Thesis Co-advisor

(Ekapol Chuangsuwanich, Ph.D.)

. . . . . Second Thesis Co-advisor

(Sira Sriswasdi, Ph.D.)

. . . . . Examiner

(Asst. Prof. Sukree Sinthupinyo, Ph.D.)

. . . . . Examiner

(Assoc. Prof. Cholwich Nattee, D.Eng.)

กรพัฒน์ ปรีชากุล: การออกแบบสถาปัตยกรรมหาตำแหน่งแบบสอนอย่างอ่อนสำหรับภาพทางการแพทย์. (REDESIGNING WEAKLY SUPERVISED LOCALIZATION ARCHITECTURES FOR MEDICAL IMAGES) อ.ที่ปรึกษาวิทยานิพนธ์หลัก : ศ. ดร. บุญเสริม กิจศิริกุล, อ.ที่ปรึกษาวิทยานิพนธ์ร่วม : อ. ดร. เอกพล ช่วงสุวนิช, อ. ดร. สิระ ศรีสวัสดิ์ 55 หน้า.

ด้วยขนาดฐานข้อมูลภาพเอ็กซ์เรย์ปอดที่ใหญ่ขึ้นทุกวัน ทำให้การใช้โมเดลเชิงลึกให้ผลที่ดีมากขึ้นและถูกใช้งานเป็นผู้ช่วยให้แก่รังสีแพทย์ในการวินิจฉัยภาพเอ็กซ์เรย์ โดยผลการทำนายประกอบกับคำอธิบายจากโมเดลนั้นสามารถช่วยรังสีแพทย์ได้ในมุมที่ว่าโมเดลเชิงลึกนั้นได้เรียนรู้มาจากฐานข้อมูลขนาดใหญ่และมีความหลากหลายอย่างยิ่ง ซึ่งคำอธิบายผลการทำนายดังกล่าวนั้นมักอยู่ในรูปแบบของแผนที่ความร้อน (heatmap) โดยจะชี้ไปยังบริเวณที่สำคัญเกี่ยวข้องกับความผิดปกติหนึ่ง ๆ ที่พบเห็นในภาพเอ็กซ์เรย์ แต่เนื่องจากฐานข้อมูลขนาดใหญ่มักไม่มีผลเฉลยถึงระดับตำแหน่ง แผนที่ความร้อนดังกล่าวจึงถูกสร้างขึ้นจากเทคนิคด้าน class-activation map (CAM) ซึ่งเป็นหนึ่งในวิธีการอธิบายโมเดล แต่ CAM นั้นมีข้อจำกัดคือแผนที่ความร้อนที่ได้นั้นจะมีความละเอียดต่ำอย่างยิ่ง และความละเอียดต่ำนี้ทำให้การอธิบายผลการทำนายต่อรังสีแพทย์นั้นทำได้ไม่ดีเท่าที่ควร ซึ่งเป็นปัญหาหลักที่วิทยานิพนธ์นี้แก้ไข ด้วยการเสนอสถาปัตยกรรมใหม่ชื่อว่า **Pyramid Localization Network** หรือย่อว่า PYLON (**ไพลอน**) ที่สามารถสร้างแผนที่ความร้อนที่มีความละเอียดสูงและมีความถูกต้องสูง โดยได้ทำการทดลองเปรียบเทียบอย่างละเอียดบนฐานข้อมูลภาพเอ็กซ์เรย์สองชุด ชื่อว่า NIH's Chest X-Ray 14 และ VinDr-CRX กับโมเดลอื่น ๆ หลายชนิดที่ถูกนำเสนอมาก่อนหน้า จากผลการทดลองสามารถสรุปได้ว่าไพลอนนั้นให้แผนที่ความร้อนที่มีความแม่นยำสูงที่สุดและสูงกว่าโมเดลอื่นอย่างมากบนหลายคลาสความผิดปกติบนทั้งสองฐานข้อมูล โดยที่ความสามารถดังกล่าวมิได้ทำให้ความแม่นยำในผลการทำนายของไพลอนลดลงแต่อย่างใด โดยยังได้เสนอวิธีการถ่ายโอนความรู้ระหว่างฐานข้อมูลที่เรียกว่า **two-phase** โดยได้ทำการทดลองถ่ายโอนความรู้จากฐานข้อมูล NIH ไปยัง VinDr-CXR ซึ่งให้ผลดีกว่าวิธีการปกติอย่างยิ่ง วิทยานิพนธ์นี้ยังได้ศึกษาขึ้นตอนของการสร้างไพลอน โดยทำการทดลองเปรียบเทียบแต่ละชิ้นส่วนของไพลอนเพื่ออธิบายถึงความสำคัญของแต่ละชิ้น และยังได้ทำการศึกษาว่าใช้ global average pooling ซึ่งถูกใช้ในสถาปัตยกรรม DeeplabV3+ และ PAN นั้นส่งผลเสียในคุณภาพของแผนที่ความร้อนอีกด้วย ซึ่งความรู้ทั้งหมดนี้ได้ถูกใช้ในการพิจารณาในขั้นตอนการสร้างไพลอนซึ่งทำให้ได้แผนที่ความร้อนที่มีคุณภาพสูงที่สุดดังได้กล่าวมาข้างต้น

ภาควิชา	คณิตศาสตร์และวิทยาการคอมพิวเตอร์	ลายมือชื่อนิสิต	.....
สาขาวิชา	วิศวกรรมคอมพิวเตอร์	ลายมือชื่ออ.ที่ปรึกษาหลัก	.....
ปีการศึกษา	2563	ลายมือชื่อ.ที่ปรึกษาร่วม	

6070106021: MAJOR COMPUTER ENGINEERING
KEYWORDS: DEEP LEARNING, MEDICAL IMAGING, CHEST X-RAY, EXPLAINABILITY, CLASS-ACTIVATION MAP, WEAKLY-SUPERVISED LEARNING

KONPAT PREECHAKUL : REDESIGNING WEAKLY SUPERVISED LOCALIZATION ARCHITECTURES FOR MEDICAL IMAGES. ADVISOR : PROF. Boonserm Kijsirikul, D.Eng., THESIS COADVISOR : Ekapol Chuangsuwanich, Ph.D., Sira Sriswasdi, Ph.D. 55 pp.

With increasing availability of large scale chest x-ray datasets, deep learning models have made great strides to improve the quality of chest x-ray readings by providing insights to radiologists from vast and diverse datasets they have trained on. These insights are the classification predictions and explainability of the predictions which are in the form of heatmaps. Due to the lacking of high-quality spatial annotation like bounding boxes in large chest x-ray datasets, the heatmaps are generated from class-activation map (CAM) methods. CAM methods however produce heatmaps with very limited resolutions which limit the usefulness of the models for radiologists. This thesis aims to alleviate this problem of limited resolution heatmaps. We propose **Pyramid Localization Network** (PYLON) which produces not only high-resolution but also high-accuracy heatmaps. We carefully demonstrated and compared PYLON in multiple datasets, namely NIH's Chest X-Ray 14 and VinDr-CXR, against many previous works showing that PYLON produced the best CAM regarding point localization accuracy while maintaining classification performance. The improvements from PYLON over the previous works are substantial in most classes in both datasets. We also propose **two-phase** fine-tuning method for transferring knowledge of PYLON across datasets, from the NIH to the VinDr-CXR, while maintaining high-level of localization accuracy. We justified the design of every component of PYLON via a series of elaborate ablation studies. We studied the negative effect of global average pooling on the accuracy of heatmaps which we demonstrated in DeeplabV3+ and PAN. These findings were important for the design of PYLON which achieved the highest quality CAM.

Department	:	Computer Engineering	Student's Signature	.....
Field of Study	:	Computer Engineering	Advisor's Signature	.....
Academic Year	:	2020	Co-advisor's signature	.....

This thesis is a product of countless unfruitful hypotheses coupled with countless disappointments. These letdowns humbled one researcher, all the while built his character to be tougher and brighter. I have learned valuable lessons through the long years of this humbling experience. I never have felt a feeling of insurmountable difficulty during the years of studentship, yet so abundantly felt during the years of scholarship. It was too easy to be frustrated as I no less was. However, one should not be so fast to conclude that "it is not worthwhile". I still think it is. Not so many would have this unique chance of "delving deeper" than it is justifiable by feeble reasons of profitability which an industry may require. Under the umbrella of profitability, many things are not worth knowing. Yet to those like us, we just *need* to know to satisfy our thirst of understanding. We are a special breed who would not settle to *half* an explanation. I am convinced that this *fire* of unease will be best quenched in the path of scholarship which I am forever grateful to have taken.

The computation used in the experiments in this thesis was partially supported by CMKL University, Thailand.

CONTENTS

LIST OF FIGURES

Chapter I

INTRODUCTION

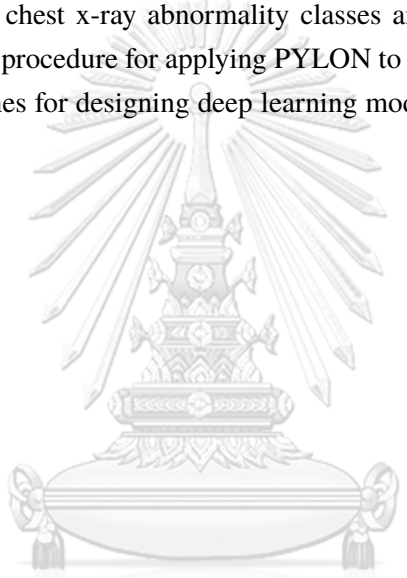
Automatic abnormality classification of chest x-rays via deep learning sees great interests in both academic and industry over the years due to increasing availability of large public chest x-ray datasets, such as NIH's Chest X-Ray 14 (Wang et al., 2017), CheXpert (Irvin et al., 2019), MIMIC-CXR (Johnson et al., 2019), Padchest (Bustos et al., 2020), and VinDR-CXR (Nguyen et al., 2020). Deep learning, known for its superior image classification performances on several natural image benchmarks (He et al., 2015; Russakovsky et al., 2015), is well suited for this medical image classification task where data are abundant and high accuracy is required for clinical adoption.

However, classification accuracy alone is not enough for healthcare application where the patient's health is at stake. In practice, deep learning models are used as second opinions alongside radiologists. Hence, the ability to convey the models' decisions in human understandable forms are highly important. One common way to describe the model's prediction for an input chest x-ray image is through a heatmap where a color highlights the important image regions that correspond to a class of interest. The most straightforward way to obtain such heatmaps is to train models with ground truth bounding boxes annotated by expert radiologists. Unfortunately, such spatial annotation requires substantial amount of effort and is not typically available in large quantities. For example, only 1% of the NIH's Chest X-Ray dataset of more than 100,000 chest x-ray images and the recently released VinDR-CXR dataset of 15,000 chest x-ray images possess this level of annotation. The VinDR-CXR dataset in particular was annotated by three different radiologists per image and could facilitate a systematic analysis of the impact of doctor-to-doctor variability on deep learning model performance.

The limited amount of chest x-rays with detailed spatial annotation necessitates another paradigm to obtain the heatmap without explicitly training the model with ground truths. A popular technique is the class-activation map (CAM) (Oquab et al., 2015; Selvaraju et al., 2017; Zhou et al., 2016), which lets us derive class-specific heatmaps as by-products of the model prediction process. CAM is a *partial* explainability method. It explains only the high-level decision of a certain class of models using a linear function which is typically easy to understand. This is achieved by producing a linear combination of the highest-level feature maps from the model for each specific class. Pioneered by Wang et al. (2017) and CheXNet (Rajpurkar et al., 2017), CAM has become a common approach for generating heatmaps for chest x-ray classification. Though there are many advances in chest x-ray classification in general (Liu et al., 2019b; Shin et al., 2016; Baltruschat et al., 2019; Zhang et al., 2020; Wang et al., 2018; Guan et al., 2018), the improvement in the quality of CAM itself is much more limited, in part, due to the lack of objective measures of the CAM's quality. Many works have proposed to use more sophis-

ticated pooling functions like MIL-pool in Li et al. (2018c) and Rozenberg et al. (2020), and LSE-pool in Wang et al. (2017) and the parameterized version LSE-LBA in Yao et al. (2018) whose architecture also produces higher resolution CAM than others.

One major limitation of CAM is its low resolution because CAM is produced by a linear combination of the *highest*-level feature map which is usually much smaller than in the input image size in deep classification networks. For chest x-ray classification, this severely hinders the ability of CAM to precisely locate small lesions such as Nodules, whose sizes are on average only 0.5% of the total image area. In this work, we propose a new architecture named Pyramid Localization Network (**PYLON**) that can generate high resolution and high accuracy CAM. We show via extensive series of experiments that PYLON improves the accuracy of CAM compared to previous works across chest x-ray abnormality classes and datasets. We also propose an effective transfer learning procedure for applying PYLON to small datasets, and we summarize requirements and guidelines for designing deep learning model capable of generating accurate CAM.



Chapter II

BACKGROUND

In this chapter, we lay grounds for readers who are not familiar with the field of deep learning in medical imaging, in particular, chest x-rays, which requires some explainability of the model. We describe some aspects of explainability methods which are directly related to medical imaging.

2.1 Deep learning in chest x-rays

The field of medical imaging has recently been affected by the advent of deep learning due to the availability of large public datasets, namely Chest X-Ray 14 (Wang et al., 2017), CheXpert (Irvin et al., 2019), MIMIC-CXR (Johnson et al., 2019), and Padchest (Bustos et al., 2020). All of the mentioned are large open datasets of chest x-ray images which are the most common type of medical imaging. Their sizes have reached to the point at which many deep learning techniques, known for their being data hungry, become practical. This is less true to varying degrees for other modes of scans, e.g. CT scans, bone scintigraphy, PET scans, and others which are available in much smaller volumes. Rajpurkar et al. (2017) showed that a deep learning model is comparable to human radiologists for pneumonia classification under a specific setting where there is only visual information available, not clinical. This is a huge milestone hence started the practical applications of deep learning models in medical imaging in both academic and industry.

2.1.1 Chest x-ray basics

Since this thesis focuses on chest x-ray, some basic knowledge about chest x-rays may be beneficial. Chest x-ray is a kind of radiograph taken by capturing x-ray emitted by a beamer through a patient chest. The varying degrees of detectable x-ray reflected on a film give a transparent look for a radiologist to read cues of abnormalities in the chest area of the patient. These abnormalities are called **findings** (abnormalities that are read from the film) and are written in a **clinical report**. Findings are not diagnoses, they are descriptive on the shape and outlook of the abnormalities, for example, *Enlarged cardiomediastinum* which says about the abnormally large size of Cardiomediastinum of the patient, but not a diagnosis of the underlying cause. Most of the times, chest x-ray is used as a screening measures for more delicate and sophisticated scans and procedures which will better determine the root cause of those abnormalities and finally diagnoses. Yet, in the medical reports, radiologists give their **impressions** on *possible* diagnoses given all the patient information known by the radiologist. The impressions could be useful information and suggestion on the next action should it be taken by the doctor of the patient who *reads* the clinical report.

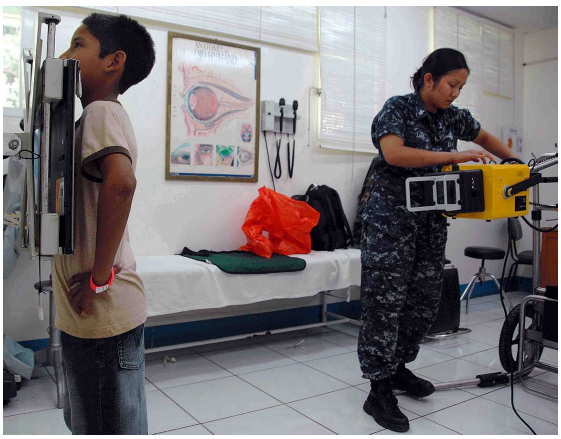


Figure 2.1: A child is taking a PA view chest x-ray. He turns his back to an x-ray beamer, his front to the film. Picture from wikipedia.org.

Figure 2.2: Example chest x-ray images from a single person. (Left) PA view. (Mid) AP view. (Right) Lateral view. The most obvious difference between PA and AP is they are the reflects of each other, but there are more subtleties which could be useful for radiologists. From MIMIC-CXR-JPG dataset (Johnson et al., 2019).

Since chest x-rays are *planar* scans (2D scans), there are always a lost of information from the projection (from a 3D chest onto a 2D film) which is alleviated to some degree by taking from different angles. The most common angles are: posterior-anterior (PA), anterior-posterior (AP), and lateral views. Figure 2.1 indicates how a PA view is taken. A person turns their *back* on an x-ray beamer and their front to the film. The opposite is true for the AP view which is usually taken when taking the PA view is not possible. PA view is considered a "frontal" view which is also the most preferred view for chest x-ray. When possible, lateral views are used as a compliment to PA and AP views, mainly for depth information which is largely ignored in PA and AP.

The difference between AP and PA is best depicted in Figure 2.2, that is they are the reflects of each other. This does not mean that both views carry the same information, for one, the heart silhouette of both views are not the same due to the different distances to the film. The subtle differences in these views are beyond the scope of this thesis.

Dataset	Total size	Report	Labels
Chest X-ray 14 (Wang et al., 2017)	108,948	No	14
Chexpert (Irvin et al., 2019)	224,316	No	14
MIMIC-CXR-JPG (Johnson et al., 2019)	377,110	Yes	14
Padchest (Bustos et al., 2020)	160,000	Yes (Spanish)	192
VinDr-CXR† (Nguyen et al., 2020)	15,000	-	15

Table 2.1: Comparing large open chest x-ray datasets. † The dataset was not constructed from reports but by hand-labelling images directly by radiologists.

2.1.2 Chest x-ray datasets

Most large scale chest x-ray datasets are *retrospective* which means they are from clinical reports written and cumulated over the years. Clinical reports are not yet ready to be consumed by deep classification models being in English full-text with considerable variance between writing styles and keyword choices of radiologists. An automatic label extraction pipeline is applied on the reports to gather only the related keywords which indicate interested abnormalities, for example Cardiomegaly and Pleural effusion. Negbio (Peng et al., 2018), CheXpert (Irvin et al., 2019) and machine learning model (Bustos et al., 2020) are possible choices for extracting the labels from full-text reports. Both Negbio and CheXpert have information extraction background. They work on the same principle which is, first, extract keywords (called mentions), then identify the mentions, whether they are positive or negative mentions. These methods involve handcrafting rules to detect the positive and negative mentions which could be quite complex in real reports, yet these extractors have been shown to have high precision and recall. Bustos et al. (2020) took a different approach to extract the labels. They designed a sequence classification model which takes in the report sentences and outputs the multi-label binary classification indicating the positive keywords. While taking less manual work in the modeling process, it requires a training dataset which needs to be manually extracted in the first place.

Understanding the process by which the ground truths are constructed in large datasets is important to understand its limitations and flaws such as there are many labellers potentially with different level of expertise and different *biases* in the labelling process which could contaminate the consistency of the labels. We include large open chest x-ray datasets in Table 2.1. Note that some datasets only release their extracted labels not their original reports for privacy reasons.

The most relevant datasets to this thesis is NIH's Chest X-ray 14 (Wang et al., 2017) and VinDr-CXR (Nguyen et al., 2020) both were selected because they contain *bounding-box* annotation on the chest x-ray images which are important for evaluation in this thesis.

2.1.3 Chest x-ray classification models

After the dataset is constructed and ready, ideally in the form of a table whose columns comprise of an image file name and corresponding label classes either positively labelled, nega-

tively labelled, or missing, one can apply deep learning models on it, and train it via binary cross entropy loss (the task is binary multi-label) with stochastic gradient descent. Probably the most well-known is CheXNet (Rajpurkar et al., 2017), a DenseNet-121 (Huang et al., 2017) model without any modification that classifies 14 thoracic diseases from chest x-ray. CheXNet is not the first to use a deep learning model in this task, there are a few before that (Wang et al., 2017; Yao et al., 2017), but CheXNet is the first to claim "radiologist-level" with elaborate performance evaluation against radiologists on Pneumonia classification earning considerable appearances. CheXNet also popularizes the use of DenseNet in chest x-rays for later works.

2.2 Deep model explainability and medical imaging

When a deep learning model is applied on other fields, it can get away without much need to *explain* its prediction results. Most of explanation of the model is used mostly by the researcher to fix and improve the model. However, in the field of medical imaging, the patient's health is at *stake*. Most models are used as second opinions alongside radiologists by providing helpful findings from their vast experience trained from large and diverse datasets. The interface between the model and the radiologist has become important requiring the model to explain its decision in a human understandable form.

Explainability of a machine learning model is a whole field with active research. Not all are suitable for explaining deep models, not all are suitable for explaining models with pixel inputs, e.g. images as ours, and not all provide intelligible explanations useful for radiologists. This crossed out many explainability techniques like SHAP (Lundberg and Lee, 2017) which is not suitable for pixel input, gradient and gradient-like attribution methods (Simonyan et al., 2013; Shrikumar et al., 2016, 2017) which may not be class-sensitive (Adebayo et al., 2018) and may not produce useful explanations for the use in medical imaging being sparse and disperse. If an explainability method is tasked with explaining "where a model looks at", most likely a model looks almost *everywhere*, and the *strengths* by which the model looks are shown to be difficult to obtain from convolution models used in image classification. Understandably, a convolution layer has fixed-weights, that is the strength is seemingly *fixed* regardless of the *content* of the image. Hence, most of the power of a convolution layer comes from the non-linearity itself which is difficult to explain when stacked into many deep layers. This suggests that gradient and gradient-like based methods are not to be pursued.

In fact, a *truthful* explainability method may not be the *goal* of explaining the decision made by a model after all. One needs to contemplate that truthfulness is not the goal but *understanding* is. The only truthful method is to look at the *weights* of the model itself, yet it is not very understandable, hence every explainability method is a loss of information and none is perfectly truthful. This steers us to the conclusion that a *good* explainability method is *domain* specific, and in the domain of medical image classification, it *may* be enough to *point* to the location of abnormalities of interest where the radiologists can take a look by themselves. Accepting that

this *pointing* explanation does not explain *how* it is abnormal, yet we are explaining this to an expert radiologist if the pointing is *specific* and accurate enough, the how may easily come up by the radiologists themselves. We conclude that not only a model predicts whether abnormalities of different kinds are present in the input image or not, but also it outputs *where* for a human radiologist to look to see such abnormalities. This can be represented as a *heatmap* where a color represents the degree to which an abnormality of a given kind is present at a location.

The most obvious way for a model to predict the *heatmap* is to train it explicitly via spatial annotations like bounding boxes or segmentation maps. Unfortunately, these annotations can only be obtain from expert radiologists whose time is expensive hence only small quantities of this level of annotation is available. Another way is to obtain the heatmap via an explainability method of a particular kind called class-activation map (CAM) (Selvaraju et al., 2017; Zhou et al., 2016; Oquab et al., 2015). CAM is a *partial* explainability method which means it does not really explain the process by which the model comes up with its decision, it only explains the *top most* layer of the model which is a linear layer and a linear layer is easy enough to explain. With some strong assumptions on the model, a heatmap is obtained by linearly *combining* the top most feature maps output from the convolution network, since the weights are class specific, the combined map is called *class-activation* map. CAM can refer to either the explainability method and its output. Examples of heatmaps generated by a CAM method under chest x-ray settings is shown in Figure 2.3. In case of Grad-CAM (Selvaraju et al., 2017), the weights do not need to be the top most layer though, it can be anywhere in the model as long as the weights can be linearly *approximated* by its gradient from the class output to that particular layer, however one needs to keep in mind that the quality of Grad-CAM is limited by the approximation quality of the gradient which we have discussed to be poor on complex models like convolution models.

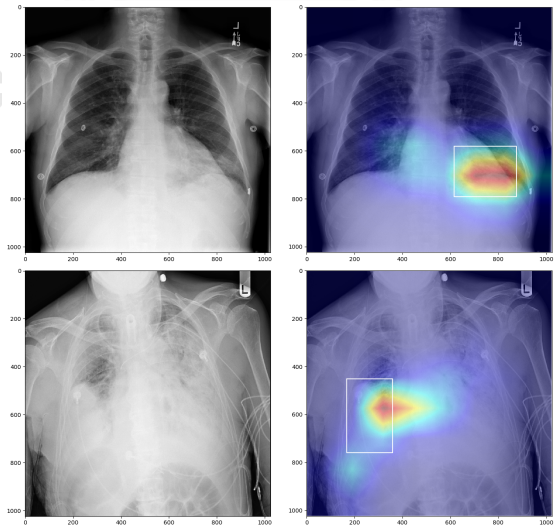


Figure 2.3: Heatmaps generated by class-activation map methods on chest x-ray images. White bounding boxes are ground truth locations of the abnormalities. Networks do not observe these bounding boxes.

2.2.1 CAM's assumptions

CAM methods allow heatmaps to be generated from a model for *free* without any spatial annotation, e.g. bounding boxes. That is training a CAM model requires only image-class labels which are abundantly available in large open datasets. The heatmaps are by-product of the process by which a model comes up with its class predictions. Yet, CAM relies on strong assumptions on the model to make sure that the heatmaps are intelligible to the beholders.

The most important assumption and the most obvious one is that the model must be **shift equivariant**. It simply means that if the model recognizes a feature at left-top location of the input, the model will output high values at the *corresponding* left-top location on its output as well. In a sense, the model is *truthful* about the location of its finding. This property is a bare minimum requirement for CAM without which there is no guarantee that the heatmap will put peaks and troughs at the right locations regarding the model and input. We can write shift equivariance property formally as:

$$f(\text{Shift}_{\Delta w,\Delta h}(\mathbf{x})) = \text{Shift}_{\Delta w,\Delta h}(f(\mathbf{x})) \tag{2.1}$$

where f is the function, and $\text{Shift}_{\Delta w,\Delta h}$ is a function that shifts its input by $\Delta w, \Delta h$. In practical terms, the function f shifts its output as its input shifted by the same amount.

It is known that a convolution layer is *likely* to be shift equivariant due to its nature of applying identically to every patch of the input. We say *likely* because when *padding* is applied to the input, the borders can be identified by the function, by looking at specific patterns such as areas of zeros. This allows a sophisticated-enough function to apply different transformations on specific parts of the input especially near borders diminishing its equivariance property. However, this is unlikely to happen in practice. As we hold a convolution layer as shift equivariant, it implies that *fully-convolution network* (Long et al., 2015) is also shift equivariant because all of its components are shift equivariant, including pooling functions. When a model breaks the property and is applied by CAM method to generate heatmaps, the results become unstable and not reliable. It may work or fail to varying degrees which destroys its credibility of an explainability method. A full-convolution network is hence important framework for building more sophisticated models that work well with CAM.

Another consideration when using CAM methods is about the *task* itself. Recall that the most important thing about model's explanation is understandability by the beholder. We first come up with an ideal vision of model's explanation, and we find an explainability method to satisfy it. In medical image classifications, CAM is reasonable in most settings with goals of heatmaps pointing to the prominent location of a class on the input. In other words, if the task is "whether an entity X in the image?" CAM tends to work as expected. When the question becomes more *complex* and the entity in question is not a patch of opacity or an explicitly pointable object on the image, i.e. a symptom or a concept may not be pointable, CAM will be hard pushed and its output might not reflect the location of the entity. This limitation comes from the fact

that CAM explains only the last linear layer of a model and makes *no attempt* to explain further into the earlier layers. If the entity requires complex reasoning between multiple cues, the reasoning is likely to be beyond the capacity of the last linear layer, and thus the reason must be done in *earlier* layers of the model which is out of scope of the CAM explainability, hence CAM is oblivious to it and the heatmap will register none of it.

2.2.2 Fully convolution network

Shift equivariance property is very important for the usability of CAM, and a specific family of deep network that satisfies the shift equivariance property is fully-convolution network. We explore more about it in details.

Fully-convolution network (FCN) was proposed in (Long et al., 2015) for semantic segmentation task. The original work has no interest regarding shift equivariance, but rather on its efficiency as a way to apply the same function on an image in a silding window fashion. Though the name was introduced with a specific network, nowadays it has been used as a *family* of networks which includes all recent classification networks like ResNet (He et al., 2016a,b), DenseNet (Huang et al., 2017), and EfficientNet (Tan and Le, 2019).

VGG (Simonyan and Zisserman, 2015) is not an FCN even though most of its layers are convolutional except a few last fully-connected layers and a *flatten* layer. The use of a flatten layer breaks the shift equivariance property because it scrambles both spatial dimensions (H, W) and the channel dimension (C) together which destroys the separation between the spatial information in (H, W) and the class information in (C). This separation is very important to the shift equivariance because in order to allow for the shift of input to result in *merely* the shift of output, the class information, which is not changed via the shift of input, must not be kept in the (H, W) dimensions. Should there be any part in the (H, W) dimensions that contains class information, that part would not *not* change according to the shift which breaks the shift equivariance property.

Flatten is one prohibitive layer to be used in FCN, yet any pooling or even global pooling layer is fine with FCN. Again, the reason is the separation of spatial information in (H, W) and the class information in (C) which a pooling layer either local or global because a pooling layer never combines across dimensions. The use of global pooling layer is found in more recent classification networks mentioned above.

2.3 Semantic segmentation models

Semantic segmentation is the name of a computer vision task where models are tasked with classifying each and every pixel in the input image into one of many possible classes. The resultant classification *maps* describe the areas of different classes which are useful applications in visual scene understanding such as autonomous vehicle.

The task has been around for many years and is deemed a fundamental problem in computer vision. Before the age of deep learning, semantic segmentation was approached by *conditional random fields*-based models such as Krähenbühl and Koltun (2011). CRF is particularly useful in this task because to be able to predict a class of a single pixel a lot of contextual information both global and local must be taken into account making the relationship between the pixel output and the input image very complex. Where the relationships are modelled explicitly with CRF, models saw much improvement, and CRF has been associated with this task ever since.

Deep learning models saw first strides in image classification task, ImageNet in particular where the improvement from deep learning models are significant and beyond previously attainable by classical machine learning models (Krizhevsky et al., 2012; Simonyan and Zisserman, 2015; He et al., 2016a). Chen et al. (2014) adopted deep learning to solve semantic segmentation with the heritage of CRF making it a hybrid model. FCN (Long et al., 2015) was among the earliest models to approach semantic segmentation task with pure deep learning. With later works along the same line, such as Unet (Ronneberger et al., 2015) and DeeplabV3 (Chen et al., 2017a), CRF has seen less and less use in semantic segmentation owing to the fact that with increasing capacity of the segmentation model the association between context and the prediction become more doable with pure deep learning model alone. Yet, some have not forgone from CRF completely. There are attempts to incorporate CRF directly into the convolution operation itself such as Su et al. (2019).

All modern semantic segmentation models make use of *lateral* connections (also known as skip connections) proposed in U-Net (Ronneberger et al., 2015). These modern models make the separation between *encoder* and *decoder* explicit. The encoder is tasked to glean the context and class-specific information from the input image and summarize them in a more compact form usable by the decoder who predicts a class of each pixel given the information. Lateral connections connect between each encoder-decoder block pair, earlier encoders connect to later decoders and later encoders connect to earlier decoders essentially forming a U-shape connection hence the name U-Net. It is common for the encoder to taper in resolution to gather information from a larger context with high efficiency, yet this reduces the feature map's resolution drastically. At the prediction time, decoder needs to *upscale* the small feature map to be close to the input resolution for pixel-level prediction, without the lateral connection this step is very hard to do and poor results ensued early deep segmentation models. The lateral connection alleviates the upscaling problem of the decoder by providing it with high-resolution *guiding* signals from the encoder to help output high resolution prediction better. Most of the models adopted this framework except DeeplabV3 (Chen et al., 2017a) and DeeplabV3+ (Chen et al., 2018) which mainly exercise atrous (or dilated) convolution which maintains high resolution feature maps all the way at the cost of more computation.

After the success in natural language processing (Bahdanau et al., 2014) and speech recog-

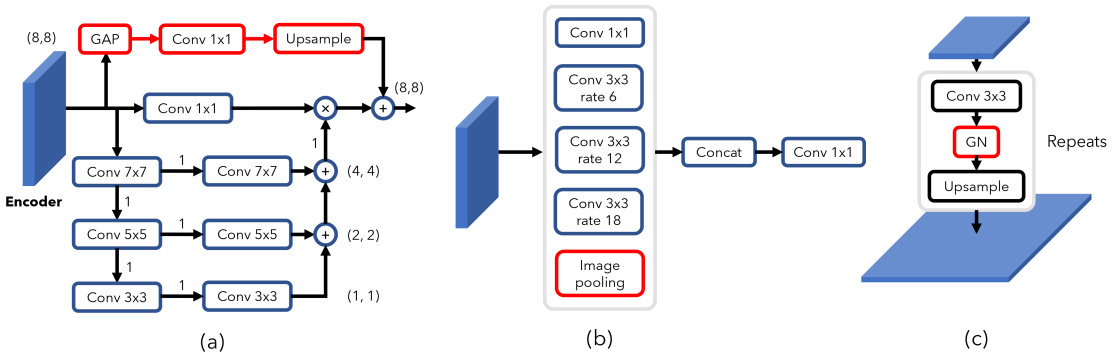


Figure 2.4: **(a)** FPA module of PAN, **(b)** ASPP module of DeeplabV3+, **(c)** A part of the semantic segmentation brach of FPN. See the original paper for more details.

nition (Chorowski et al., 2015), attention mechanisms have been adopted in image classification in a form of *channel-wise* attention in SENet (Hu et al., 2018). A more generalized attention is *spatial-attention* seen in Chen et al. (2017b). Attention mechanisms also found uses in semantic segmentation as well as in Pyramid Attention Network (PAN) (Li et al., 2018a) in its FPA module where the spatial-attention is adopted in a form of pyramidal network shown in Figure 2.4 **(a)** seen as hanging rungs.

DeeplabV3+ (Chen et al., 2018) and PAN (Li et al., 2018a) see the use of *bias* connection shown in *red* in Figure 2.4 **(a, b)**. We call these biases because their functions are additive signals, and they apply uniformly across spatial dimensions, that is they involve global pooling and then upsampling back to the original dimensions. This mechanism potentially makes the model train faster by easing the work by "putting solid colors on the canvas" so that the work left is just putting in the details.

FPN (Kirillov et al., 2019) is another well known architecture which was proposed to be applicable to many related vision tasks such as object detection, semantic segmentation, instance segmentation and panoptic segmentation. The model is considered not complicated and fast with reasonable performance which explains its wide spread use. FPN for semantic segmentation uses group norm (Wu and He, 2018) instead of the usual batch norm (Ioffe and Szegedy, 2015) in its segmentation branches shown in Figure 2.4 **(c)**. The name FPN has its root from "pyramid" network which involves a long line of research on a deep network that works well on input of different scales and sizes which is crucial in object detection and semantic segmentation (Lin et al., 2017; Chen et al., 2017a; Kirillov et al., 2019; Li et al., 2018a; Chen et al., 2018). Notable characteristics of a pyramid network are having lateral connections and having prediction heads on different level of feature maps.

Chapter III

RELATED WORKS

Class activation map (CAM). CAM is a *partial* explainability method trying to explain only the high-level decision of the model where only linear functions are involved making it relatively straightforward to explain. The word "CAM" may refer to a family of explainability method or heatmaps generated by a CAM method itself. Oquab et al. (2015) was the first to generate heatmaps via CAM from a deep classification model with global max pooling. The model learns with image classification labels and gets heatmaps as a by-product for free as intermediate outcome before the pooling for predictions. Zhou et al. (2016) popularized CAM from a global average pooling model. Later, Selvaraju et al. (2017) proposed GradCAM which is a CAM method for more general models not just an image classification model. GradCAM can be applied to an image-to-sequence model, e.g. image captioning, by approximating the weights from the output gradients, potentially from the output of an LSTM layer, with respect to the feature maps of the convolution network. One needs to keep in mind that GradCAM relies on the quality of the gradient which becomes less reliable when the model through which the gradient is calculated is increasingly complex.

CAM's limitations. Heatmaps generated by CAM are marked by only focusing on small regions of the object, for example, CAM may focus only on the face of a cat not its entire body. This behavior depends on the global pooling functions to some degree, for example global max pooling is known the generate the most *focused* heatmaps (Oquab et al., 2015), following by log-sum-exp pooling (Wang et al., 2017), then global average pooling (Zhou et al., 2016). The seemingly overly focused heatmaps are *expected* behavior though due to the fact that for most objects only a small part of them is *discriminative* enough to rely on for classification, hence the classifier does not care much about the rest. Another well-known problem of CAM is its inability to address *all* occurrences of the same kind of object (Bae et al., 2020). For example, CAM can only reliably focus on *one* of potentially many instances of objects of the same class. It can do better than that but there is no guarantee. This problem is also expected from the way we train a classification model. The training labels are just whether "there is" or "there is not" an object of this class in the image. The meaning of "there is" does not imply how many, a single occurrence would suffice. If a classifier can identify one instance of the object, there is no further training signal to push the classifier anymore.

CAM's improvments. The overly focusing problem and the single instance problem of CAM stem from the same root which is that CAM only cares the most *discriminative* part of the image. If one could *force* CAM to look at other parts, CAM might put some weights there and has more coverage. Solving this problem usually involves *erasing* parts of the input image or the feature maps. It has been shown that Cutout (DeVries and Taylor, 2017) and DropBlock

(Ghiasi et al., 2018) ameliorate the problem to some degree by making the discriminative parts of the image less reliable for the model to look at. The erasing can be made more aggressive by removing those parts marked as *important* by the CAM itself which gives CAM even stronger urges to learn other less discriminative parts. This is also essentially the same line of work as *adversarial erasing* which could use more sophisticated pipeline to train the model with selective erasing parts of its input (Choe and Shim, 2019; Li et al., 2018b; Fan et al., 2017; Zhang et al., 2018b; Singh and Lee, 2017; Ren et al., 2020). It is worth noting that erasing methods are usually not robust to parameter settings, datasets, and classifier models. Object size priors can be directly enforced on the feature maps **?** to prevent the overly small focusing regions given that one knows the object size priors. Object boundary priors can also be enforced as shown in Zhu et al. (2017); Ahn et al. (2019); Huang et al. (2018), though not likely to be very useful in the case of chest x-ray.

CAM for weakly-supervised object localization. From the beginning, CAM can already generate reasonably accurate heatmaps in terms of localization, that is it can *point* to the location of the object precisely, though cannot cover the whole object. It is possible that with further improvements CAM can be used to learn an object detector without any bounding box supervision which should drastically reduce the cost of annotation. Notable works along this line tries to make use of many tricks notably *adversarial erasing* mentioned before. Ren et al. (2020) uses adversarial erasing coupled with teacher-student training and pseudo bounding box generation. Interestingly, Huang et al. (2020) shows that though CAM focuses on a small region of the image, after a small change on the image via augmentation, CAM tends to focus on a different location of the image. More comprehensive coverage of heatmaps can be obtained by multiple forward passes on the same image under different augmentations. The union of which is used as guiding signal to the CAM itself to produce more coverage heatmaps.

Unsupervised quality measurement of CAM. Can heatmaps generated by CAM be trusted? If spatial annotations like bounding boxes are available, they can be used to verify the CAM, but what if they are not available? Most works opted for qualitative assessment delegating the job to the eyes of the beholders. Assuming that CAM shows the regions *important* for the decision, in a sense that without these regions class predictions would change, the quantitative assessment of the quality of CAM can be done by *erasing* particular regions suggested by CAM, and the change in *confidence* of predictions reflect the correctness of CAM. A perfect CAM is the one when whose regions are removed from the input then the confidence goes to zero (Lin et al., 2019). The same idea can work in *reverse* to obtain an explainability heatmap.

CAM in chest x-rays. When deep models are used in health related applications, much care is given to make sure that the models are doing exactly what they are supposed to do. In chest x-ray classification, models are expected to explain their decisions to the radiologists. It now becomes a norm for chest x-ray models to use CAM for explanation as was pioneered by Wang et al. (2017), CheXNet (Rajpurkar et al., 2017) and CheXpert (Irvin et al., 2019). Li et al.

(2018c) proposed to use the multi-instance learning principle (MIL) to learn better localization on chest x-rays. CAM has also been applied to CT scan models (Li et al., 2020). In chest x-ray report generation task, CAM can be obtained from the attention weights in LSTM report generator (Liu et al., 2019a). Due to the limitations of CAM, their heatmaps are of low resolution resulting in not very accurate localization of the diseases.

Black-box explainability method. Black-box methods assume no particular knowledge about the inner workings of the model. The explainability method accesses the model only via inputs and outputs. One particular example is SHAP (Lundberg and Lee, 2017) which adjusts the input of the model principally to get the average contribution of each feature, pixels in our case, yet it has some limitations on highly correlated set of features making it not suitable to explain image classification models (Frye et al., 2019). This has an interesting connection to *erasing* explainability techniques where parts of the input are removed and the feature importance is the magnitude of change in prediction confidence Zolna et al. (2020); Fong and Vedaldi (2017); Chang et al. (2018).

Gradient attribution explainability method. Earliest explainability methods for deep convolution network was of the gradient attribution family where the *attribution* (importance) of each pixel is *propagated* backward from the prediction. The discontinuous of gradient is a usual problem with networks with ReLU activation rendering unpleasing saliency maps (analogous to heatmaps). Many attempts proposed gradient-substitute with better continuous property (Shrikumar et al., 2017; Bach et al., 2015). Some heuristics like winner-take-all are also applied instead of gradient propagation for more artistically pleasing results (Zhang et al., 2018a). It has been shown by Adebayo et al. (2018) that gradient attribution methods may not be class specific at all which diminishes their use as explainability methods. Zeiler and Fergus (2014) used transposed weights to construct the prominent patterns in the input image. Simonyan et al. (2013) proposed to alter the input image by the gradient signals to amplify the qualities-related to a specific class, in a sense, creating an exemplar of a class by amplifying class-related features.

Semantic segmentation models. Traditionally, semantic segmentation was approached by CRF-based models such as Krähenbühl and Koltun (2011). Chen et al. (2014) adopted deep learning to solve semantic segmentation with the heritage of CRF making it a hybrid model. FCN (Long et al., 2015) was among the earliest models to approach semantic segmentation task with pure deep learning. Lateral skip connection has been proposed in Unet (Ronneberger et al., 2015) and is found useful in most succeeding models. PAN (Li et al., 2018a) sees adoption of channel-wise attention and spatial-attention. Atrous convolution is used almost exclusively in DeeplabV3 (Chen et al., 2017a) and DeeplabV3+ (Chen et al., 2018) instead of lateral skip connection. FPN (Kirillov et al., 2019) is another well known architecture which was proposed to be applicable to many related vision tasks such as object detection, semantic segmentation, instance segmentation and panoptic segmentation.

Chapter IV

PYLON: A DEEP NETWORK FOR HIGH-RESOLUTION AND HIGH ACCURACY CLASS-ACTIVATION MAP

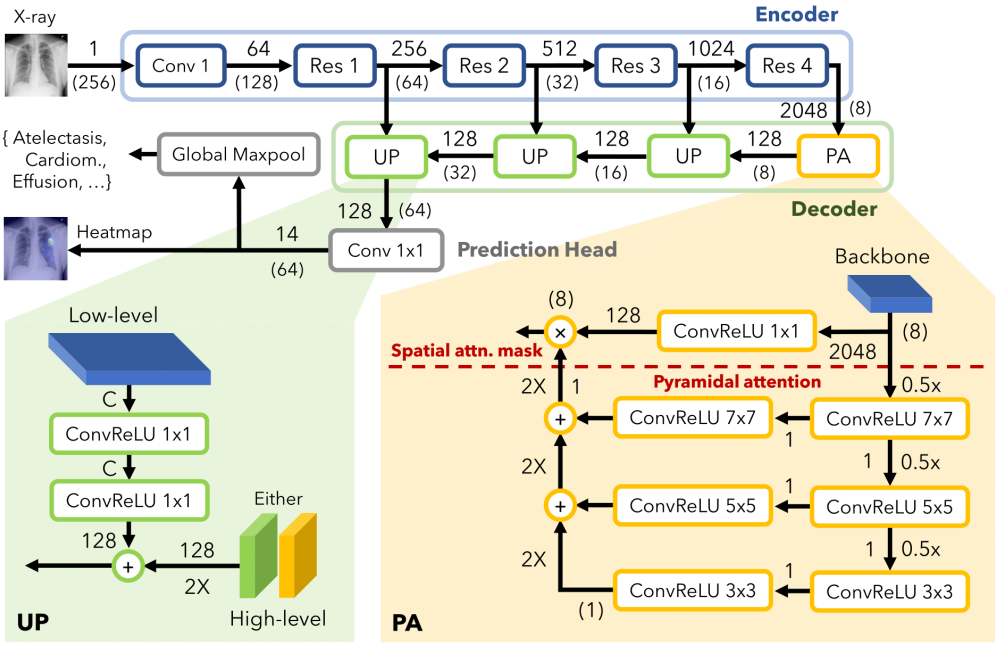


Figure 4.1: **Pyramid Localization Network (PYLON)** with its **Pyramid Attention (PA)** and **Upsampling (UP)** modules. The model consists of *three* parts: an encoder, a decoder, and prediction head. The encoder could be ResNet, DenseNet or others. Here we assume the input of size 256×256 and ResNet-50 as the encoder. **Heatmap** is the CAM output. **Global Maxpool** is used to turn class heatmaps into classsification predictions. **2X** refers to bilinear upsampling. **0.5X** refers to 2×2 max pooling. Each **ConvReLU** is a convolution layer followed by batch normalization and ReLU activation. The numbers (along the arrows) denote the number of *channels* while the numbers in *parentheses* denote the *size* of the feature map. In **PA**, there is a *pyramidal attention* path that produces a *spatial attention mask* (has one channel) which multiplies with the main Conv 1×1 path as a spatial attention mechanism.

Deep models with explainable decisions are important in medical imaging where the patient's health is at stake. As the models would be used as a second opinion alongside human radiologists, ideally explanation should be understandable and useful for them. Most works in chest x-ray proposed to use heatmaps as the means of explanation which are generated from CAM methods. Heatmaps are easily understandable in a sense that they contain only *where* not *how* which could be thought of as lacking delicate information. Yet, they are useful for radiologists as second opinions knowing that radiologists are expert themselves. By looking at the right place, they could themselves come up with the *how*. Though CAM can be obtained almost for free from image-class labels which are widely available, heatmaps generated from a typical deep classification model have very limited resolutions. The limitation stems from the architecture of the classification model itself which has low resolution top-layer feature maps. In this chapter, we propose a model to alleviate this limitation.

Though improvements on CAM can be made from many angles, we argue that the most

obvious one (and also the largest gain) could be obtained by increasing feature map's resolution of the model which, in turn, will increase the resultant heatmaps produced by CAM. We designed **PYLON** specifically to produce high resolution and high accuracy heatmaps. PYLON has *three* parts: an encoder, a decoder, and a prediction head. The encoder can be any of the well-known deep architectures for image classification such as ResNet (He et al., 2016a), DenseNet (Huang et al., 2017), VGG (Long et al., 2015), and EfficientNet (Tan and Le, 2019) and their corresponding variants to name a few. The decoder is composed of two different modules namely **Pyramid Attention module (PA)** and **Upsampling module (UP)**. The encoder, decoder, and prediction head collectively are called PYLON which stands for Pyramid Localization Network. The overall architecture is shown in Figure 4.1. In PYLON, there is only *one* PA module but *many* UP modules, each upscales the input signal by 4 times ($2\times$ horizontally and $2\times$ vertically) with the help of a lateral skip connection from a corresponding encoder's *block*. There can be as many UPs as the encoder's blocks *minus* one. A *block* of layers in this context is a group of layers which operate on the same resolution. In ResNet and its variants, there are 4 total *blocks*.

Pyramid attention module (PA) is a form of *spatial-attention* which is found useful in image classification (Hu et al., 2018; Chen et al., 2017b) and semantic segmentation (Li et al., 2018a). The word "pyramid" in its name has its root in a long line of research which utilizes different input scales in order to handle inputs of varying sizes especially in object detection and semantic segmentation (Lin et al., 2017; Chen et al., 2017a; Kirillov et al., 2019; Li et al., 2018a; Chen et al., 2018). As shown in Figure 4.1 **(PA)**, **Pyramidal attention** in the PA module compresses its input signal into a *single* channel while consecutively *downscaled* it into varying degrees. At each scale, the signal is passed through a non-linear transformation and *upscaled* back for a final combination (across scales) into a **spatial attention mask** (which has one channel). The spatial attention mask is multiplied to the transformed encoder feature maps like a typical attention mask. PA takes the upper most feature maps of the encoder and outputs feature maps of the same dimensions but with fewer channels called *decoded feature maps*. The decoded feature maps will be upscaled by the following UPs modules.

Upsampling module (UP) is a lightweight two-layer 1×1 convolution with batch norm (Ioffe and Szegedy, 2015) and ReLU activation of the lateral skip connection from a corresponding encoder's block. The output from each layer of this is *added* to the bilinearly *upscaled* ($2\times$ horizontally and $2\times$ vertically) *decoded feature maps* from the PA module or the previous UP module.

Prediction head. After upscaled by the last UP module, the *decoded feature maps* are combined by a linear function (equivalent to Conv 1×1) for each specific class to get a corresponding CAM for each class. To train with classification loss, each class's CAM is *globally max pooled* to get each class prediction. The training of PYLON is the same as training a normal classification network except that high accuracy CAMs are produced as a by-product.

PYLON is simple and fast by nature relying mostly on Conv 1×1 with little computation and memory overhead. The success of PYLON is to delegate work on its encoder which has seen continual improvements (Simonyan and Zisserman, 2015; Krizhevsky et al., 2012; He et al., 2016a,b; Huang et al., 2017; Hu et al., 2018; Tan and Le, 2019).

Effective *two-phase* transfer learning approach

Transfer learning has become a standard practice to train neural nets on more limited datasets. When the accuracy of CAM is of concern, we propose a **two-phase** fine-tuning approach. The process begins by fine-tuning the decoder and the prediction head while *freezing* the encoder until convergence. Then, we unfreeze the encoder and train the whole network together until convergence again. Both phases are trained with the same learning rate. This two-phase fine-tuning approach can retain much better localization knowledge from the source dataset.

Code availability

We provide codes for reproducing the main results and extra qualitative results in https://github.com/cmb-chula/pylon.

Chapter V

RESULTS

5.1 Datasets

This thesis conducted experiments on *two* datasets: NIH's Chest X-Ray 14 (Wang et al., 2017) and VinDr-CXR (Nguyen et al., 2020). Both of which contain region-level annotation which is required for comparing class-activation map (CAM) accuracy. The **NIH's Chest X-Ray 14** dataset is larger with more than 100,000 frontal (AP and PA) chest x-ray images mostly annotated by extracting related 14 chest abnormalities automatically from clinical reports. 880 images (across 8 abnormalities) of all are manually annotated with bounding boxes by a board certified radiologist. **VinDr-CXR** is smaller with 18,000 PA chest x-ray images (15,000 available at the time of writing). All of which are annotated with bounding boxes across 15 abnormalities by three different radiologists with at least 8 years of experience (from a pool of 17 radiologists). Note that the official dataset has 28 abnormalities however not yet released at the time of this study. We used the Kaggle competition version of this dataset. In this study, *no bounding box was used* to train a deep network, only used for evaluation.

Metric on the accuracy of CAM

This study used a metric called **point localization accuracy** which is the ratio of how frequently a heatmap produced by CAM *points* within the region of ground truth bounding boxes given that an abnormality is present. The *point* is obtained by finding the maximum confidence location (row and column) on the heatmap (upscaled to have the same dimension as the input). This metric is selected to ensure a fair comparison between heatmaps produced from different models which can have different resolutions and specificity levels (in terms of the size of highlighted areas). Moreover, the ground truth annotations are *rectangular* bounding boxes while the actual regions of abnormalities are not necessarily rectangular. Therefore, we cannot assume the whole region of the bounding box to be correct. Thus, a fair comparison regardless of the specificity of heatmaps under such limitation is comparing only the *point* of highest confidence on the heatmap.

Previous works also used metrics like intersection over proposed bounding box (IoBB, also known as IoR) which requires *two* thresholds to be selected; one is the prediction threshold (usually 0.5), another is the intersection threshold which varies across different works ranging 0.1 to 0.5. Wang et al. (2017) coupled IoBB with IoU (intersection over union) requiring a model's prediction to meet *either* of IoBB or IoU intersection thresholds to get a score. The downside of this kind of metric is its sensitivity to thresholds selection which might be vary across different models of different natures. To accommodate comparison with other studies, we occasionally used IoBB as a metric.

Metric on classification performance

Alongside the point localization accuracy, which assumed that an abnormality is already present, we also employed *area-under receiver operating characteristic curve* (AUROC) to measure the classification performance in terms of sensitivity and specificity for each abnormality class regardless of the classification thresholds.

5.2 Benchmark models

We compare our proposed model PYLON against a series of baseline models and previous works. We always use the same ResNet-50 encoder to ensure fair comparisons. The **Baseline** model is a common CAM pipeline with global max pooling (Oquab et al., 2015). Note that Baseline is also comparable to GradCAM (Selvaraju et al., 2017) which approximates the high-level decision process as a linear function. In the Baseline, the high-level decision process is already linear hence no approximation and GradCAM is equivalent to the traditional CAM. **Li2018** (Li et al., 2018c) proposes multi-instance loss function (MIL) and utilizes a semi-supervised learning technique with partial bounding box annotation and is a common benchmark model in many works (Yao et al., 2018; Rozenberg et al., 2020). Since there are only limited number of works that proposed image classification model with high resolution CAM, we also include strong image segmentation models including **Unet** (Ronneberger et al., 2015), **FPN** (Kirillov et al., 2019) with batch normalization (**BN**), **PAN** (Li et al., 2018a) and **DeeplabV3+** (Chen et al., 2018), which might produce high accuracy CAM due to their naturally high resolution outputs. Each segmentation-based baseline model was *turned* into an image classification model by adding a global max pooling on top of its output to get a class prediction score. It is worth noting that the performance of Unet depended hugely on its decoder's number of channels. We selected (256, 128, 64, 64, 64) as the number of channels which is fairly larger than the default of (256, 128, 64, 32, 16).

There are other related works that we cannot compare directly. **Wang 2017** (Wang et al., 2017) proposes a typical CAM model (Zhou et al., 2016) coupled with log-sum-exp pooling (LSE pool) which should already be represented by our Baseline model. **Yao 2018** (Yao et al., 2018) proposed a model with high resolution CAM with parameterized LSE pool called LSE-LBA pool, however due to limited details provided, we could not re-implement the results claimed in the paper. For these two models, we will compare them with their reported numbers. Rozenberg et al. (2020) proposed a more numerically stable MIL loss function based on Li et al. (2018c) with some architectural improvements (Zhang, 2019; Su et al., 2019). However, they focused on semi-supervised training with bounding box supervision and did not report numbers without bounding box supervision hence we will not compare with their results.

5.3 Experimental details

We tried our best to control for the variances between model settings when possible. Each experiment was rerun five times using different initializations to estimate the variances of model performances. Confidence intervals were calculated with Student's T distribution (with n=5). All experiments used Adam (Kingma and Ba, 2015) with learning rate 10^{-4} with no weight decay. The learning rate is reduced by $5\times$ when the loss on the *validation* dataset does not improve for *two* consecutive epochs. The experiment was stopped when the learning rate was reduced by more than *two* times. The *best* checkpoint was selected based on the loss on the validation set of the respective dataset. These *best* checkpoints were evaluated on the test set of the respective dataset. All images were resized to 256×256 (or 512×512) before feeding to models. All models used ResNet-50 with Imagenet pretrained weights at their cores to reduce the possible variances. All experiments were run with mixed-precision floating point. **Augmentation.** Random rotation up to 90 degrees, random horizontal flip, random contrast and brightness in range (0.5, 1.5), random crop with random size in range (0.7, 1.0) and random aspect ratio from 4:3 to 3:4. **Transfer learning and fine-tuning.** The standard approach is to start the training from pretrained weights without freezing any part of the model using the aforementioned set of hyperparameters.

5.4 Performance on NIH's Chest X-Ray 14

Name	Weighted avg.	Macro avg.	Atelectasis	Cardiom.	Effusion	Infiltration	Mass	Nodule	Pneumonia	Pneumoth.
(Avg. area)	-	-	0.03	0.18	0.07	0.10	0.04	0.01	0.10	0.05
Baseline	0.46 ± 0.03	0.44 ± 0.03	0.31 ± 0.04	0.99 ± 0.01	0.32 ± 0.08	0.60 ± 0.03	0.42 ± 0.04	0.09 ± 0.04	0.62 ± 0.08	0.16 ± 0.02
Li 2018	0.49 ± 0.01	0.46 ± 0.01	0.36 ± 0.04	0.96 ± 0.02	0.50 ± 0.05	0.60 ± 0.05	0.51 ± 0.02	0.06 ± 0.02	0.55 ± 0.05	0.18 ± 0.03
Unet	0.45 ± 0.04	0.44 ± 0.04	0.24 ± 0.13	0.77 ± 0.28	0.39 ± 0.08	0.59 ± 0.09	0.60 ± 0.10	0.15 ± 0.04	0.61 ± 0.05	0.16 ± 0.03
FPN (BN)	0.53 ± 0.02	0.50 ± 0.02	0.38 ± 0.07	$\mathbf{1.00} \pm 0.01$	0.42 ± 0.08	0.63 ± 0.06	0.59 ± 0.07	0.14 ± 0.09	$\mathbf{0.71} \pm 0.03$	0.14 ± 0.02
DeeplabV3+	0.45 ± 0.05	0.43 ± 0.04	0.26 ± 0.07	0.81 ± 0.30	0.50 ± 0.07	0.56 ± 0.10	0.50 ± 0.04	0.07 ± 0.04	0.53 ± 0.07	$\mathbf{0.22} \pm 0.06$
PAN	0.38 ± 0.17	0.37 ± 0.16	0.24 ± 0.29	0.51 ± 0.57	$\mathbf{0.63} \pm 0.09$	0.35 ± 0.41	0.63 ± 0.04	0.10 ± 0.20	0.30 ± 0.32	0.18 ± 0.08
PYLON (ours)	$\mathbf{0.62} \pm 0.01$	$\mathbf{0.60} \pm 0.01$	$\mathbf{0.50} \pm 0.02$	0.99 ± 0.02	0.54 ± 0.05	$\mathbf{0.71} \pm 0.03$	$\mathbf{0.67} \pm 0.04$	$\mathbf{0.48} \pm 0.06$	$\mathbf{0.71} \pm 0.02$	0.20 ± 0.03

Table 5.1: Point localization accuracy on **NIH's** Chest X-Ray 14 dataset with the input image size of **256**. Accuracies are reported alongside their 95% confidence interval (n=5).

NIH's Chest X-Ray14 (Wang et al., 2017) has been the main dataset for evaluating CAM accuracy the previous works (Wang et al., 2017; Li et al., 2018c; Yao et al., 2018; Rozenberg et al., 2020; Liu et al., 2019b). The original train/test split of this dataset is available and was used in our experiments. Our validation set was obtained by splitting from the original train split. The final split has the ratios between train/validation/test of 70:7:23.

The results are summarized in Table 5.1 which shows that CAM produced by PYLON can capture the location of abnormality in chest x-ray images more accurately than previous techniques for most of the classes by large margins (95% confidence interval is provided). The largest improvements compared to the second best model were found in Atelectasis (0.5 vs 0.38 point localization accuracy) and Nodule (0.48 vs 0.15 point localization accuracy). Figure 5.1 illustrates the high accuracy and *specificity* of CAMs produced by PYLON compared to other models which tend to highlight larger areas of the chest x-ray images. Note that FPN (BN)

produced CAMs of the same resolution as those of PYLON while being much poorer in accuracy and specificity. This suggests that while high resolution CAMs are important, they are not equally accurate.

After comparing PYLON against **Wang 2017** (Wang et al., 2017) and **Yao 2018** (Yao et al., 2018) in Table 5.2, we see that PYLON's CAMs are much more accurate than Yao 2018 in both classes reported at the same resolution. Even at the disadvantage of lower resolution, PYLON was still better than Yao 2018, albeit with smaller margins, and much better than Wang 2017 in all classes, who uses a much larger image size.

Name	Resolution	Atelectasis	Cardiom.	Effusion	Infiltration	Mass	Nodule	Pneumonia	Pneumoth.
IoBB > 0.5 or IoU > 0.5									
Wang 2017	1024	0.28	0.87	0.33	0.42	0.14	0.01	0.38	0.18
Yao 2018	512	-	-	-	0.53	-	0.22	0.35	-
Baseline	512	0.38 ± 0.04	0.97 ± 0.02	0.51 ± 0.04	0.61 ± 0.03	0.48 ± 0.03	0.07 ± 0.04	0.69 ± 0.04	0.20 ± 0.04
PYLON (ours)	512	$\mathbf{0.48} \pm 0.03$	0.98 ± 0.02	$\mathbf{0.54} \pm 0.03$	$\mathbf{0.67} \pm 0.02$	$\mathbf{0.69} \pm 0.05$	$\mathbf{0.42} \pm 0.08$	$\mathbf{0.74} \pm 0.02$	$\mathbf{0.20} \pm 0.04$
PYLON (ours)	256	0.37 ± 0.04	$\mathbf{1.00} \pm 0.01$	0.42 ± 0.06	0.60 ± 0.06	0.59 ± 0.06	0.23 ± 0.03	0.67 ± 0.05	0.16 ± 0.02

Table 5.2: Comparing with Wang et al. (2017) and Yao et al. (2018) on the **NIH's** dataset. Yao 2018's method is sensitive to hyperparameter selection, we reported the best numbers (across multiple hyperparameters). CAM's accuracy depends hugely on the input resolution. For a fair comparison, this should be a controlled variable. Numbers are reported alongside their 95% confidence intervals where available.

As raw chest x-ray images are much larger than the input size of deep classification models, we further explored the impact of increasing input size from 256×256 to 512×512 on the quality of CAM. Table 5.3 shows that increasing input size substantially improves the point localization accuracy of CAM for all models, with PYLON still achieving the highest overall performance. While most models still struggled in Nodule class, PYLON excelled in it with 0.55 vs 0.34 point localization accuracy of the second best. Example CAMs are provided in Figure 5.2

Name	Weighted avg.	Macro avg.	Atelectasis	Cardiom.	Effusion	Infiltration	Mass	Nodule	Pneumonia	Pneumoth.
Baseline	0.58 ± 0.01	0.55 ± 0.01	0.51 ± 0.03	0.95 ± 0.03	0.58 ± 0.04	$\mathbf{0.70} \pm 0.04$	0.60 ± 0.01	0.19 ± 0.07	0.70 ± 0.03	0.20 ± 0.02
Li 2018	0.53 ± 0.04	0.51 ± 0.04	0.38 ± 0.06	0.94 ± 0.04	0.43 ± 0.06	0.67 ± 0.04	0.48 ± 0.12	0.23 ± 0.07	0.59 ± 0.02	$\mathbf{0.39} \pm 0.03$
FPN (BN)	0.59 ± 0.03	0.57 ± 0.03	0.45 ± 0.06	0.97 ± 0.02	0.60 ± 0.06	0.67 ± 0.02	0.61 ± 0.05	0.34 ± 0.18	0.70 ± 0.04	0.21 ± 0.04
PYLON (ours)	$\mathbf{0.65} \pm 0.02$	$\mathbf{0.64} \pm 0.02$	$\mathbf{0.56} \pm 0.04$	$\mathbf{0.98} \pm 0.02$	$\mathbf{0.61} \pm 0.05$	0.69 ± 0.06	$\mathbf{0.74} \pm 0.05$	$\mathbf{0.55} \pm 0.03$	$\mathbf{0.74} \pm 0.02$	0.21 ± 0.05

Table 5.3: Point localization accuracy on **NIH's** Chest X-Ray 14 dataset with the input image size of **512**. Accuracies are reported alongside their 95% confidence interval (n=5).

Finally, on the classification side, most of the classification ability of a model architecture depends on its encoder in this case ResNet-50 which is the same across all models. Table 5.4 shows that all models except Li 2018 have similar classification performance. The difference in classification accuracy though marginal likely came from the difference in pooling layers. While others used the same global max pooling, Li 2018 used their multi-instance pooling.

5.5 Performance on VinDr-CXR

VinDr-CXR (Nguyen et al., 2020) was a newly proposed dataset with 18,000 posterior-anterior (PA) chest x-ray images. We did not have the official test dataset, hence we split the

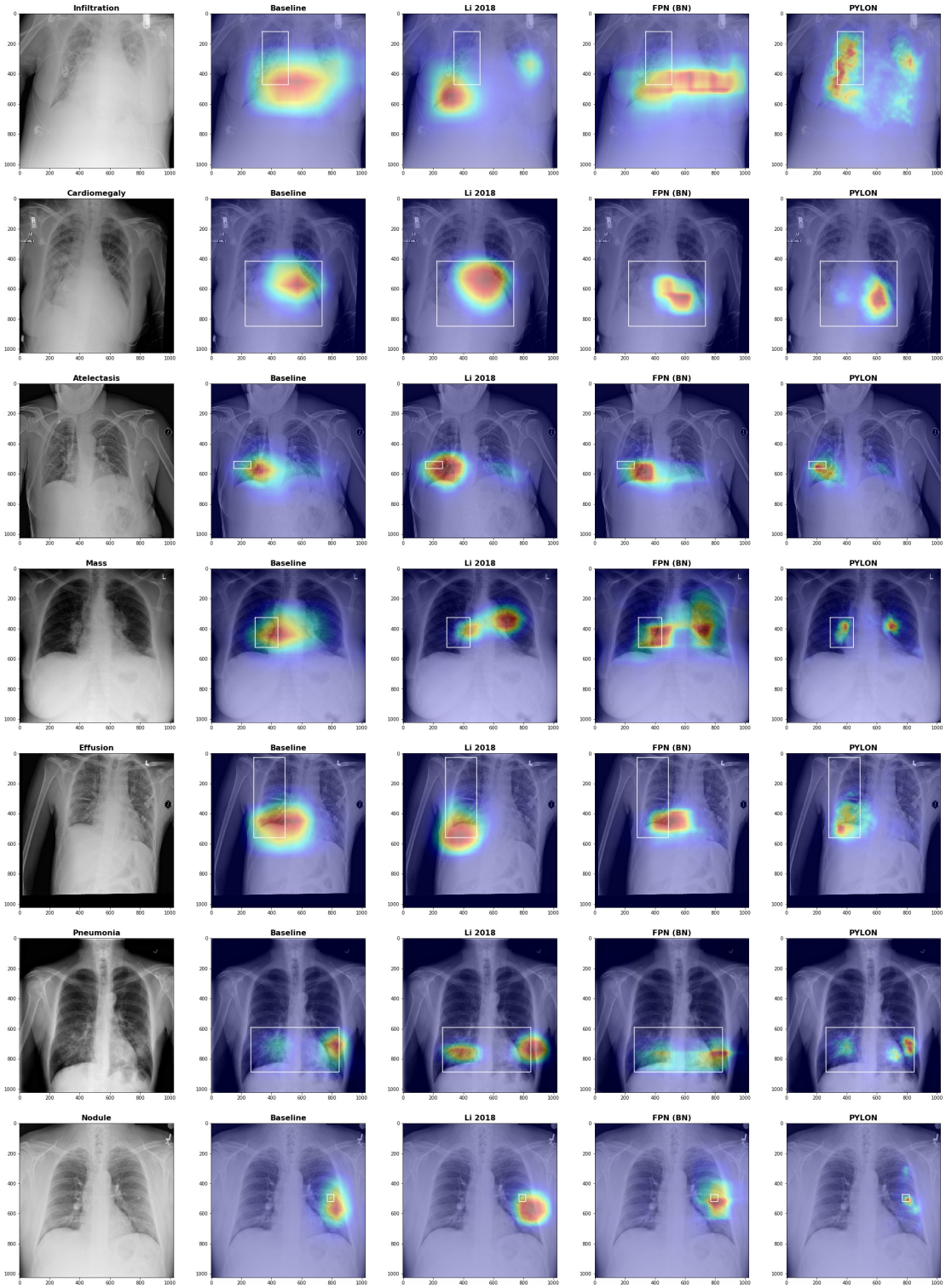


Figure 5.1: Examples of CAM on the **NIH's** dataset. Best viewed in colors.

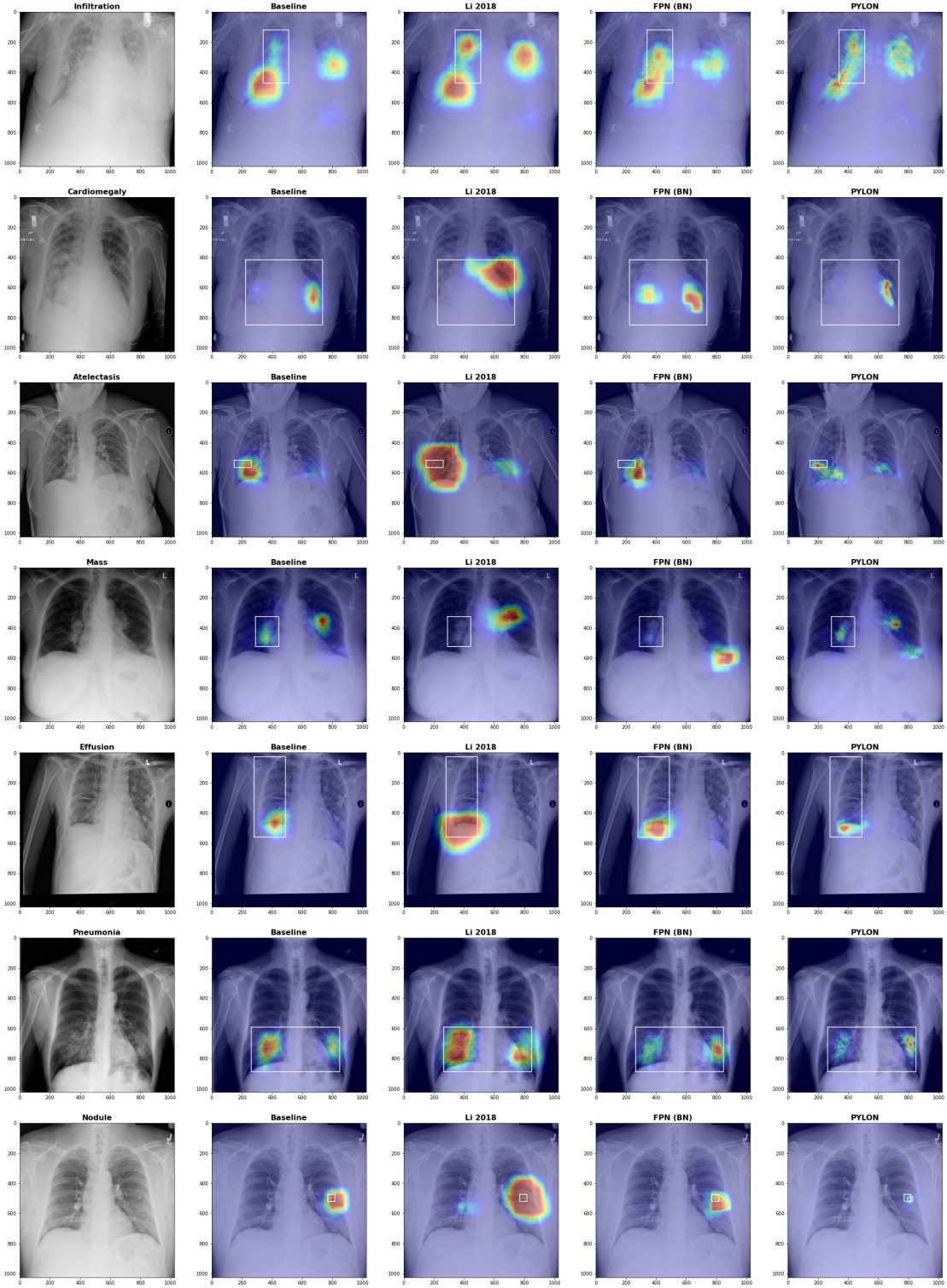


Figure 5.2: Examples of CAM from models with on the **NIH's** dataset with **512** input size. Best viewed in colors.

Name	Input resolution	Weighted avg.	Macro avg.
Baseline	256	0.795 ± 0.001	0.819 ± 0.002
Li 2018	256	**0.796** ± 0.002	**0.821** ± 0.002
FPN (BN)	256	0.795 ± 0.003	0.817 ± 0.005
PYLON (ours)	256	0.794 ± 0.001	0.819 ± 0.002
Baseline	512	0.797 ± 0.001	0.821 ± 0.002
Li 2018	512	**0.804** ± 0.001	**0.829** ± 0.002
FPN (BN)	512	0.799 ± 0.001	0.821 ± 0.003
PYLON (ours)	512	0.797 ± 0.002	0.819 ± 0.005

Table 5.4: AUROC results on **NIH's** Chest X-Ray 14 dataset. Numbers are reported alongside their 95% confidence interval (n=5).

Class	(Avg. area)	Without transfer learning				With transfer learning		
		Baseline	**Li 2018**	**FPN (BN)**	**PYLON**	**Baseline**	**PYLON**	**PYLON** (two-phase)
Weighted Avg.	-	0.29 ± 0.02	0.25 ± 0.05	0.35 ± 0.05	**0.37** ± 0.02	0.32 ± 0.03	0.46 ± 0.08	**0.55** ± 0.05
Macro Avg.	-	0.31 ± 0.02	0.29 ± 0.04	0.36 ± 0.03	**0.42** ± 0.03	0.34 ± 0.02	0.48 ± 0.04	**0.55** ± 0.03
Aortic enlargement	0.02	0.15 ± 0.08	0.24 ± 0.15	**0.33** ± 0.24	0.19 ± 0.17	0.20 ± 0.14	0.44 ± 0.39	**0.63** ± 0.20
Atelectasis	0.06	0.36 ± 0.05	0.35 ± 0.06	0.40 ± 0.09	**0.45** ± 0.14	0.47 ± 0.09	0.55 ± 0.10	**0.64** ± 0.07
Calcification	0.02	0.10 ± 0.04	0.12 ± 0.05	0.13 ± 0.02	**0.24** ± 0.17	0.13 ± 0.04	0.34 ± 0.16	**0.39** ± 0.12
Cardiomegaly	0.07	**0.76** ± 0.11	0.34 ± 0.14	0.70 ± 0.15	0.68 ± 0.31	0.78 ± 0.09	**0.83** ± 0.16	0.82 ± 0.06
Consolidation	0.05	0.56 ± 0.05	0.56 ± 0.07	0.61 ± 0.05	**0.77** ± 0.03	0.51 ± 0.04	0.79 ± 0.03	**0.83** ± 0.06
ILD	0.15	0.54 ± 0.05	0.62 ± 0.07	0.62 ± 0.13	**0.72** ± 0.04	0.60 ± 0.07	0.71 ± 0.05	**0.76** ± 0.08
Infiltration	0.06	0.42 ± 0.10	0.41 ± 0.05	0.56 ± 0.11	**0.71** ± 0.04	0.48 ± 0.06	**0.76** ± 0.03	0.74 ± 0.05
Lung Opacity	0.05	0.36 ± 0.04	0.42 ± 0.05	0.42 ± 0.05	**0.55** ± 0.03	0.38 ± 0.03	0.60 ± 0.07	**0.66** ± 0.03
Nodule/Mass	0.02	0.17 ± 0.06	0.15 ± 0.05	0.20 ± 0.03	**0.32** ± 0.07	0.23 ± 0.04	0.35 ± 0.16	**0.44** ± 0.03
Other lesion	0.05	**0.15** ± 0.03	0.13 ± 0.02	0.13 ± 0.05	0.14 ± 0.04	0.17 ± 0.01	0.16 ± 0.04	**0.24** ± 0.02
Pleural effusion	0.05	0.34 ± 0.03	0.18 ± 0.09	0.35 ± 0.04	**0.39** ± 0.06	0.35 ± 0.03	0.34 ± 0.03	**0.43** ± 0.05
Pleural thickening	0.01	0.02 ± 0.02	**0.06** ± 0.02	0.03 ± 0.01	0.03 ± 0.02	0.02 ± 0.01	0.03 ± 0.02	**0.13** ± 0.04
Pneumothorax	0.10	0.25 ± 0.10	**0.29** ± 0.09	0.29 ± 0.09	0.25 ± 0.20	0.26 ± 0.07	0.33 ± 0.16	**0.36** ± 0.07
Pulmonary fibrosis	0.04	0.18 ± 0.07	0.22 ± 0.06	0.29 ± 0.07	**0.47** ± 0.06	0.24 ± 0.04	0.50 ± 0.10	**0.61** ± 0.02

Table 5.5: Point localization accuracy on **VinDr-CXR** dataset with the image size of **256**. Numbers are reported alongside their 95% confidence interval (n=5).

official train 15,000 images into *our* train, val, and test with ratios 70:10:20, respectively. We could not guarantee that the same patient was only in the same split because the patient IDs were not available at the time of this study. Each image was labelled by *three* different radiologists. To deal with the label differences, we created a *union* of all bounding boxes from all radiologists into one big region per class per image as the ground truth region (which may not be a rectangle anymore).

In Table 5.5, we compared the accuracy of CAM produced by **Baseline**, **Li 2018**, **FPN** (with batch normalization), and **PYLON** (ours). It shows that PYLON is considerably better than the second best, FPN (BN), in terms of macro average CAM accuracy, 0.42 vs 0.36, and PYLON still came on top for 11 of 15 classes. The classification results are compared in Table 5.6 showing the similar performance which is also the case for the NIH's experiments. The improvement in CAM's accuracy from PYLON is *less* pronounced when compared to that of the NIH's dataset. We hypothesize that this is due to the limited size of the dataset itself. We defer the verification of this hypothesis to the next section on transfer learning. Example CAMs are provided in Figure 5.3.

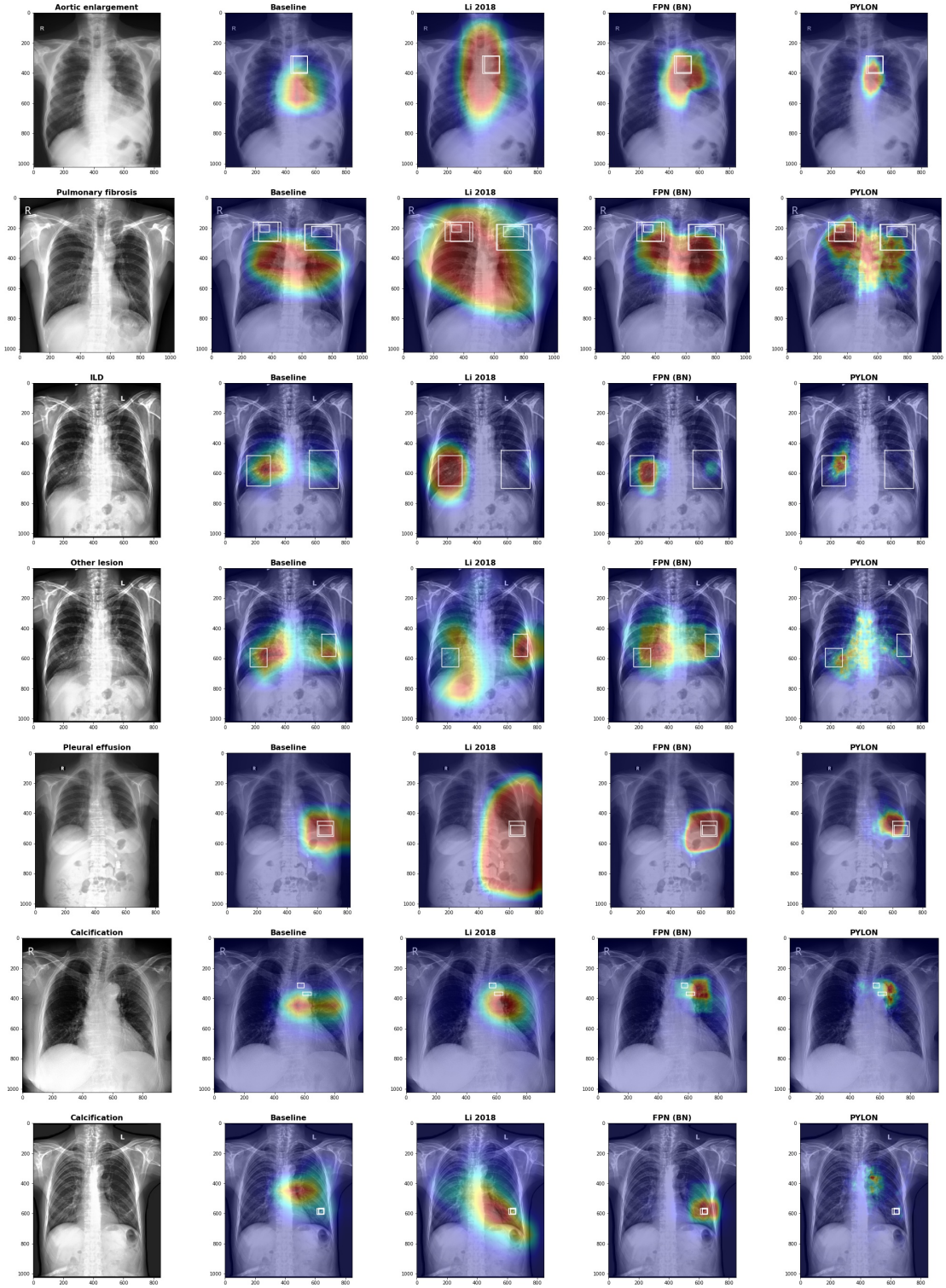


Figure 5.3: Examples of CAM from models the **VinDr-CXR** dataset. Best viewed in colors.

Name	Input resolution	Weighted avg.	Macro avg.
Baseline	256	0.970 ± 0.002	0.953 ± 0.003
Li 2018	256	$\mathbf{0.971} \pm 0.002$	$\mathbf{0.956} \pm 0.001$
FPN (BN)	256	$\mathbf{0.971} \pm 0.002$	0.954 ± 0.002
PYLON (ours)	256	0.970 ± 0.001	0.953 ± 0.001
Baseline (fine-tune)	256	$\mathbf{0.973} \pm 0.001$	$\mathbf{0.958} \pm 0.002$
PYLON (fine-tuned)	256	0.972 ± 0.002	0.957 ± 0.003
PYLON (two-step)	256	0.972 ± 0.001	0.957 ± 0.002

Table 5.6: AUROC results on **VinDr-CXR** dataset. Numbers are reported alongside their 95% confidence interval (n=5).

5.6 Transfer learning

We compared the effectiveness of the proposed *two-phase* fine-tuning approach against a standard fine-tuning approach. We included the Baseline model and its fine-tuned results to give a better perspective on the results shown in Table 5.5. Fine-tuning from pretrained weights from NIH's dataset improved CAM accuracies substantially in both Baseline model and PYLON. The proposed **two-phase** fine-tuning method further improved substantially over the standard fine-tuning approach, 0.55 vs 0.46 of the weighted average CAM accuracy, and 0.55 vs 0.37 when compared with non-pretrained PYLON. The large improvements from fine-tuning suggest that the VinDr-CXR dataset is too small for a model to learn accurate CAM from scratch. In terms of classification performance, the improvements from fine-tuning is marginal as shown in Table 5.6, yet the results show that our proposed two-phase fine tuning while improving CAM accuracy does not impede the classification performance when compared with the standard fine-tuning. Example CAMs are provided in Figure 5.4.

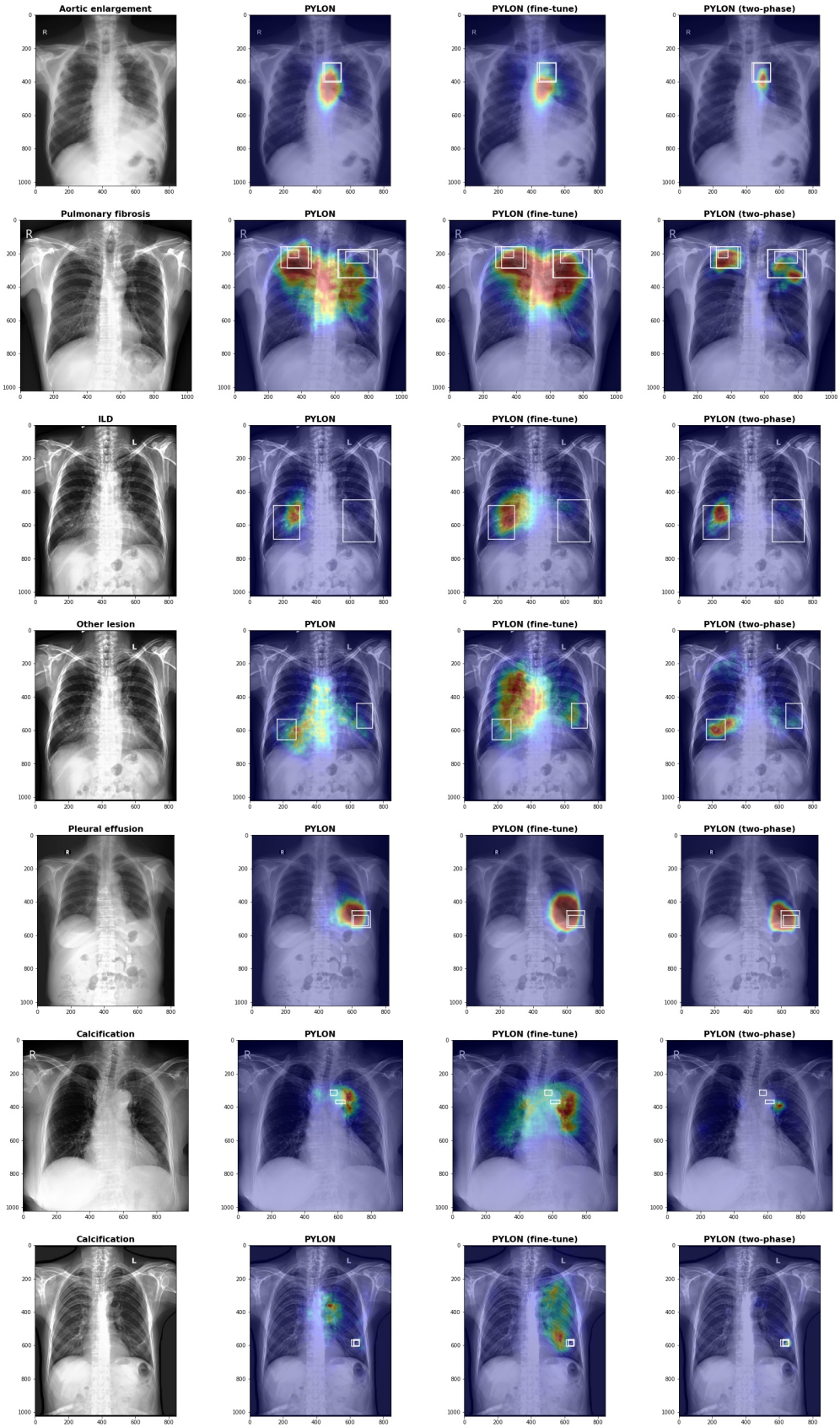


Figure 5.4: Examples of CAM from models with NIH's pretrained weights on the **VinDr-CXR** dataset. Best viewed in colors.

Chapter VI

REQUIREMENTS FOR ACCURATE CAM

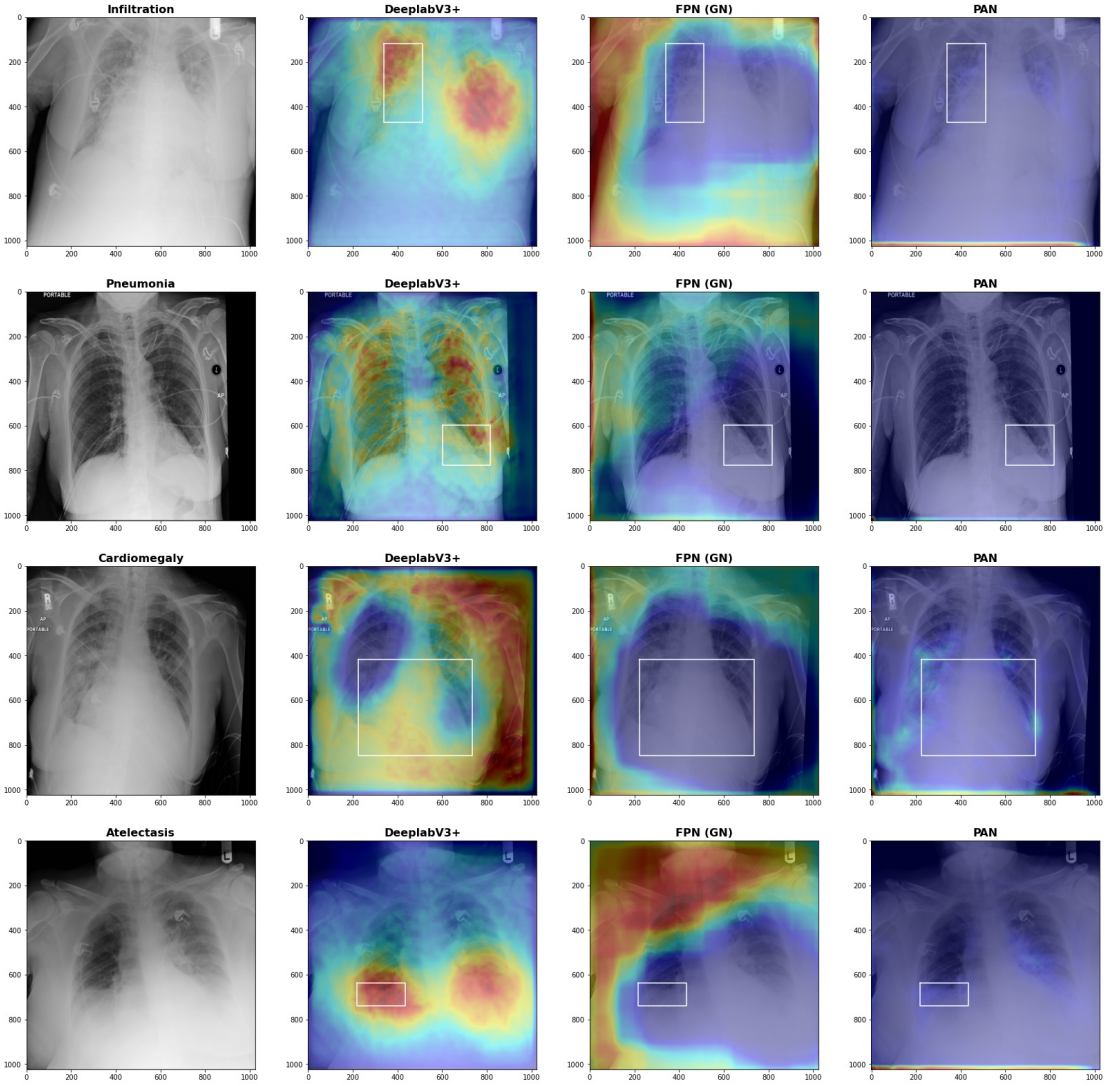


Figure 6.1: Example of poor quality heatmaps generated from DeeplabV3+, FPN, and PAN. Both FPN and PAN generated unintelligible heatmaps, though those of PAN still worked in some other classes not shown in this figure.

DeeplabV3+ (Chen et al., 2018), FPN (Kirillov et al., 2019) and PAN (Li et al., 2018a) often generated suboptimal CAM's heatmaps, examples shown in Figure 6.1. The DeeplabV3+'s heatmaps seem most intelligible though their heatmaps are not very *specific* in a sense that most of the areas are covered in colors. It is reasonable to assume that heatmaps with less specific coloring are less useful as explainability means. The original FPN's (with group norm) heatmaps are completely *unintelligible*, though the network still achieved high classification performance. Their heatmaps do not seem to correlate to any finding on the image at all and do not seem to be explaining anything. At first glance, the PAN's heatmaps seem curiously blank. This is *not*

due to the error of heatmap generation. The heatmaps instead *collapse* its colors to the borders, in this case at the bottoms of the images. It is worth noting that the collapse happens randomly from class to class and from random seed to random seed. These are unexpected results given that these models were at their times very competitive in semantic segmentation (their original task). It suggests that CAM does not work *out of the box* on all models, some underlying criteria must be met before one can apply CAM methods effectively. This is the theme of this chapter where we investigate the underlying criteria.

6.1 On FPN with group norm

The behavior of heatmaps generated from FPN with group normalization (Wu and He, 2018) proved to be elusive to understand. Knowing that normalization layers are *shift equivariant*, hence not likely to cause negative effects on CAM. The only clear reason is the use of group normalization itself because the problem was gone after substituting it with batch normalization. We have tried extensively with different group sizes to no avail. The problem also does not seem to relate to the mixed-precision training that we used as default. However, we shall take into account that the theoretical understanding of normalization layers, including group norm and batch norm, are still poor, hypotheses have been proposed and disproved over the years, and there is no simple way to fully grasp the extent of normalization's behaviors during training of deep neural networks (Ioffe and Szegedy, 2015; Santurkar et al., 2018; Arora et al., 2018; Bjorck et al., 2018). Our bottom line is that this is an *open* question suggesting further investigation.

6.2 On DeeplabV3+ and PAN

Name	Weighted avg.	Macro avg.	Atelectasis	Cardiom.	Effusion	Infiltration	Mass	Nodule	Pneumonia	Pneumoth.
(A) PAN										
PAN	$0.38 \pm \underline{0.17}$	$0.37 \pm \underline{0.16}$	$0.24 \pm \underline{0.29}$	$0.51 \pm \underline{0.57}$	0.63 ± 0.09	$0.35 \pm \underline{0.41}$	0.63 ± 0.04	$0.10 \pm \underline{0.20}$	$0.30 \pm \underline{0.32}$	0.18 ± 0.08
- GAP in FPA	0.61 ± 0.04	0.59 ± 0.03	$0.50 \pm \underline{0.10}$	1.00 ± 0.01	$0.52 \pm \underline{0.12}$	0.71 ± 0.03	0.67 ± 0.03	0.45 ± 0.04	0.69 ± 0.05	0.19 ± 0.05
- GAP in GAU	$0.50 \pm \underline{0.24}$	$0.49 \pm \underline{0.22}$	$0.38 \pm \underline{0.27}$	$0.73 \pm \underline{0.52}$	0.61 ± 0.05	$0.55 \pm \underline{0.38}$	0.64 ± 0.06	$0.20 \pm \underline{0.23}$	$0.59 \pm \underline{0.41}$	0.20 ± 0.06
- all GAP	0.61 ± 0.03	0.59 ± 0.03	0.51 ± 0.05	0.99 ± 0.02	0.53 ± 0.06	0.71 ± 0.03	0.67 ± 0.07	$0.42 \pm \underline{0.11}$	0.72 ± 0.05	0.19 ± 0.03
(B) PYLON (UP1)										
PYLON (UP1)	0.62 ± 0.01	0.60 ± 0.01	0.51 ± 0.04	0.99 ± 0.01	0.55 ± 0.06	0.72 ± 0.06	0.69 ± 0.02	0.46 ± 0.03	0.72 ± 0.05	0.18 ± 0.02
+ GAP in PA	0.60 ± 0.04	0.59 ± 0.03	0.52 ± 0.09	0.96 ± 0.05	0.57 ± 0.05	0.68 ± 0.06	0.71 ± 0.05	0.40 ± 0.05	0.69 ± 0.03	0.17 ± 0.05
+ GAP in UP	0.57 ± 0.08	0.55 ± 0.07	0.52 ± 0.05	$0.82 \pm \underline{0.45}$	0.51 ± 0.08	0.67 ± 0.05	0.65 ± 0.08	0.35 ± 0.09	0.68 ± 0.04	0.19 ± 0.02
(C) DeeplabV3+										
DeeplabV3+	0.45 ± 0.05	0.43 ± 0.04	0.26 ± 0.07	$0.81 \pm \underline{0.30}$	0.50 ± 0.07	$0.56 \pm \underline{0.10}$	0.50 ± 0.04	0.07 ± 0.04	0.53 ± 0.07	0.22 ± 0.06
- all GAP	0.44 ± 0.04	0.43 ± 0.04	0.14 ± 0.08	0.98 ± 0.02	$0.33 \pm \underline{0.13}$	0.62 ± 0.09	0.51 ± 0.06	0.06 ± 0.05	0.61 ± 0.10	0.17 ± 0.05

Table 6.1: Instability of PAN and DeeplabV3+ who have GAP in their decoders on the **NIH**'s dataset. High variance results are underlined. Numbers are reported alongside their 95% confidence interval (n=5).

The results from the experiments on segmentation networks like PAN and DeeplabV3+ saw large differences in outcomes between multiple experiments of the same settings, i.e. different random seeds. This symptom, however, did not portray itself in FPN (batch norm) which does not use any GAP. The cause of this instability is the same for *both* network architectures being the use of *global average pooling* (GAP) in their decoders. By removing such layers from each decoder, we saw immediate improvements of *stability* across experiments as shown in Table 6.1. When GAP was removed from *either* **PAN**'s FPA module (analogus to PYLON's PA module) or PAN's GAU module (analogous to PYLON's UP module), its CAM accuracy

improved drastically with reduced variances (Table 6.1 **(A)**), however there was no further improvement from removing *all* GAPs in PAN. We observed the reduction in CAM accuracy with increasing variances when *adding* GAPs into PA and UP modules in PYLON (**UP1**) (Table 6.1 **(B)**). We conclude that *any* use of GAP in PAN-like decoder lead to undesirable outcomes. In case of **DeeplabV3+**, we observed high variance in Cardiomegaly class in particular which was alleviated by the removal of GAP in its decoder (Table 6.1 **(C)**). We have shown that the use of GAP in a decoder lead to worse CAM accuracy and/or higher variance outcomes in both cases of DeeplabV3+ and PAN-like networks. We hypothesize further that this also applies to other architectures as well. This finding is the basis on why PYLON particularly refrained from using any GAP in its decoder.

6.3 PYLON's Ablation studies

Name	Point localization acc. Weighted avg.	Point localization acc. Macro avg.	Classification AUROC Weighted	Classification AUROC Macro
(A) PA module				
PYLON (UP1)	**0.62** $\pm$ 0.01	**0.60** $\pm$ 0.01	**0.794** $\pm$ 0.002	0.817 $\pm$ 0.002
PYLON (no PY, UP1)	0.60 $\pm$ 0.02	0.59 $\pm$ 0.02	0.793 $\pm$ 0.002	0.817 $\pm$ 0.002
(B) UP module				
PYLON (UP1)	**0.62** $\pm$ 0.01	**0.60** $\pm$ 0.01	**0.794** $\pm$ 0.002	0.817 $\pm$ 0.002
PYLON (Conv 3x3)	0.59 $\pm$ 0.02	0.58 $\pm$ 0.02	0.794 $\pm$ 0.002	0.818 $\pm$ 0.004
PYLON	**0.62** $\pm$ 0.01	**0.60** $\pm$ 0.01	**0.794** $\pm$ 0.001	**0.819** $\pm$ 0.002
(C) Prediction head				
PYLON (UP1)	**0.62** $\pm$ 0.01	**0.60** $\pm$ 0.01	0.794 $\pm$ 0.002	0.817 $\pm$ 0.002
PYLON (Head 3x3)	0.57 $\pm$ 0.02	0.56 $\pm$ 0.02	0.794 $\pm$ 0.001	**0.819** $\pm$ 0.001

Table 6.2: Ablation results of PYLON on the **NIH**'s dataset. Numbers are reported alongside their 95% confidence interval (n=5).

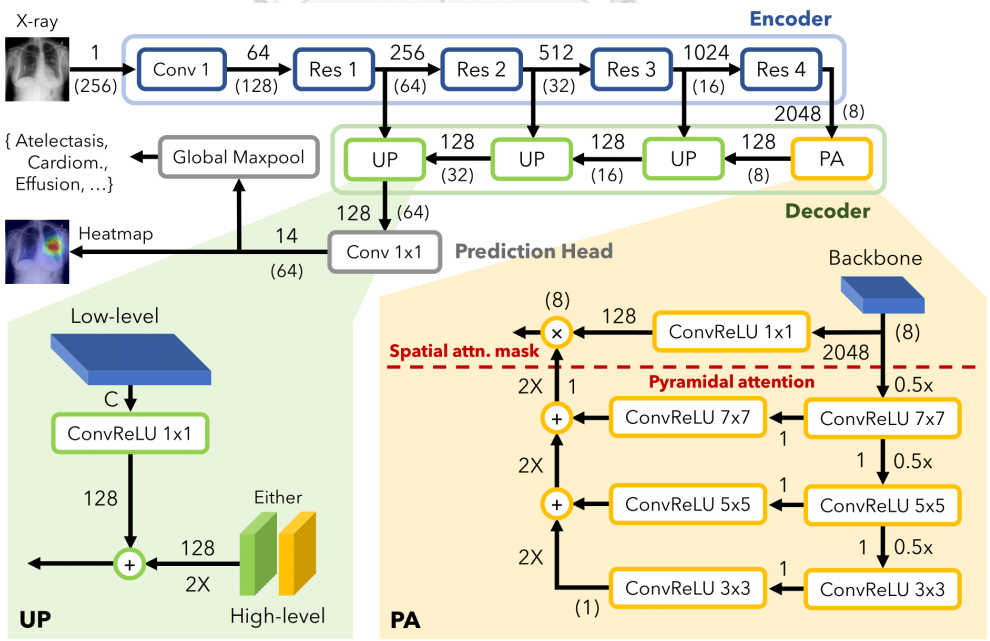


Figure 6.2: **PYLON (UP1)** variant which was used as the base model for the ablation studies. The only difference from the proposed PYLON is in its PA module. UP1 uses a *single-layer* Conv 1×1 in contrast to the original PYLON who uses a two-layer Conv 1×1.

The goal of ablation studies is to quantify the contribution of each component in PYLON

in terms of both point localization accuracy and AUROC. Our ablation studies are based on a PYLON variant called **UP1** depicted in Figure 6.2. PYLON (UP1) has a *single* Conv 1×1 in its UP module, instead of the two Conv 1×1 as in the proposed PYLON.

On the Pyramid attention (PA) module. The use of pyramidal-attention in the PA module may raise questions about its effectiveness quantitatively. We compared PYLON *without* (**no PY**) against *with* pyramidal attention (**UP1**). In Table 6.2 **(A)**, the results show that pyramidal attention improves CAM's weighted average accuracy from 0.6 to 0.62.

On the Upsampling (UP) module. We compared a single-layer Conv 1×1 variant (**UP1**), a two-layer Conv 1×1 variant (**PYLON**), and a larger Conv 3×3 variant (**Conv 3x3**). The results shown in Table 6.2 **(B)** suggest that both Conv 1×1 variants produced more accurate CAM than the Conv 3×3 variant, while the two-layer Conv 1×1 performed marginally better than single-layer Conv 1×1 in terms of classification performance, 0.819 vs 0.817 in macro AUROC.

On the prediction head. We compared the proposed prediction head, a single-layer Conv 1×1 (**UP1**) without any activation, with a larger Conv 3×3 (**Conv 3x3**). The results in Table 6.2 **(C)** show that Conv 1×1 was better regarding localization, 0.62 vs 0.57 on weighted average CAM accuracy, while being marginally worse on classification, 0.817 vs 0.819 on macro AUROC. Note that with the original PYLON (not UP1 which was used in this experiment) there was no drop in classification performance as it reached the same 0.819 macro AUROC in Table 6.2 **(B)**.

Given the limitations of the ablation studies, we observed that the smaller Conv 1×1 consistently outperformed the larger Conv 3×3 in terms of CAM's accuracy. While the larger Conv 3×3 is usually adopted in the decoders of semantic segmentation models, we advice against using it in CAM models. Though the larger Conv 3×3 might be useful for fine edge prediction, which is preferable to the semantic segmentation task, it is not useful for precise localization required for accurate CAM.

Chapter VII

UNDERSTANDING PYLON

In this chapter, we attempt to understand better *how* PYLON constructs each heatmap. We will look at the PYLON from two different perspectives: *layer-wise* perspective where we inspect what each component in PYLON outputs and how they combine to get the final heatmap, and *channel-wise* perspective where we look at each channel before they are combined to the final heatmap. We show case by a chest x-ray with Nodule class from the NIH's dataset.

7.1 Pyramid attention module (PA)

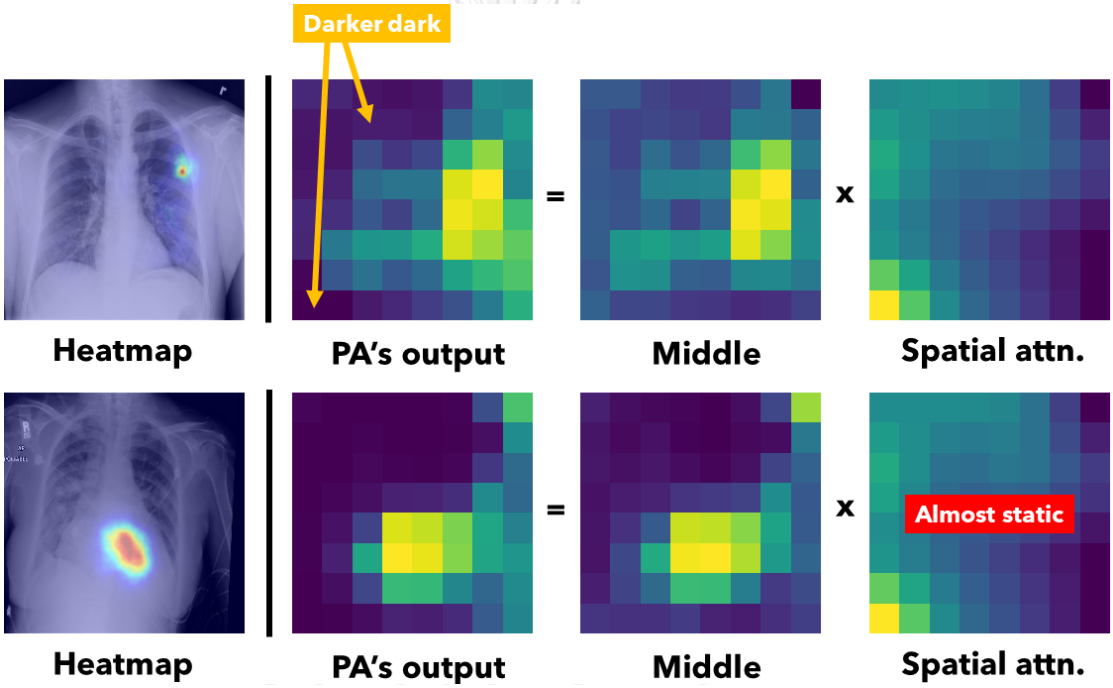


Figure 7.1: Visualizing the outputs of 1×1 Conv and pyramidal attention in the PA module from two example chest radiographs.

Pyramid attention (PA) module was shown to be useful in the ablation study (Table 6.2). However, it is unclear what does the pyramidal attention in the PA module really do to result in a gain in CAM accuracy. We visualized the outputs of both the middle path (the output of 1×1 Conv) and the pyramidal attention map in Figure 7.1. It was generally observed that the spatial attention map has a specific pattern that when multiplied with the output from the 1×1 Conv resulting in more *focused* signals (the darker part of the heatmap becomes even darker). Interestingly, the attention map was visually identical to another chest x-ray with a different abnormality. This suggests that the pyramidal attention may be context dependent, yet it is still important to the localization performance. Further studies may substituting the module with a simpler function form while still maintaining the effectiveness.

7.2 Layer-wise perspective

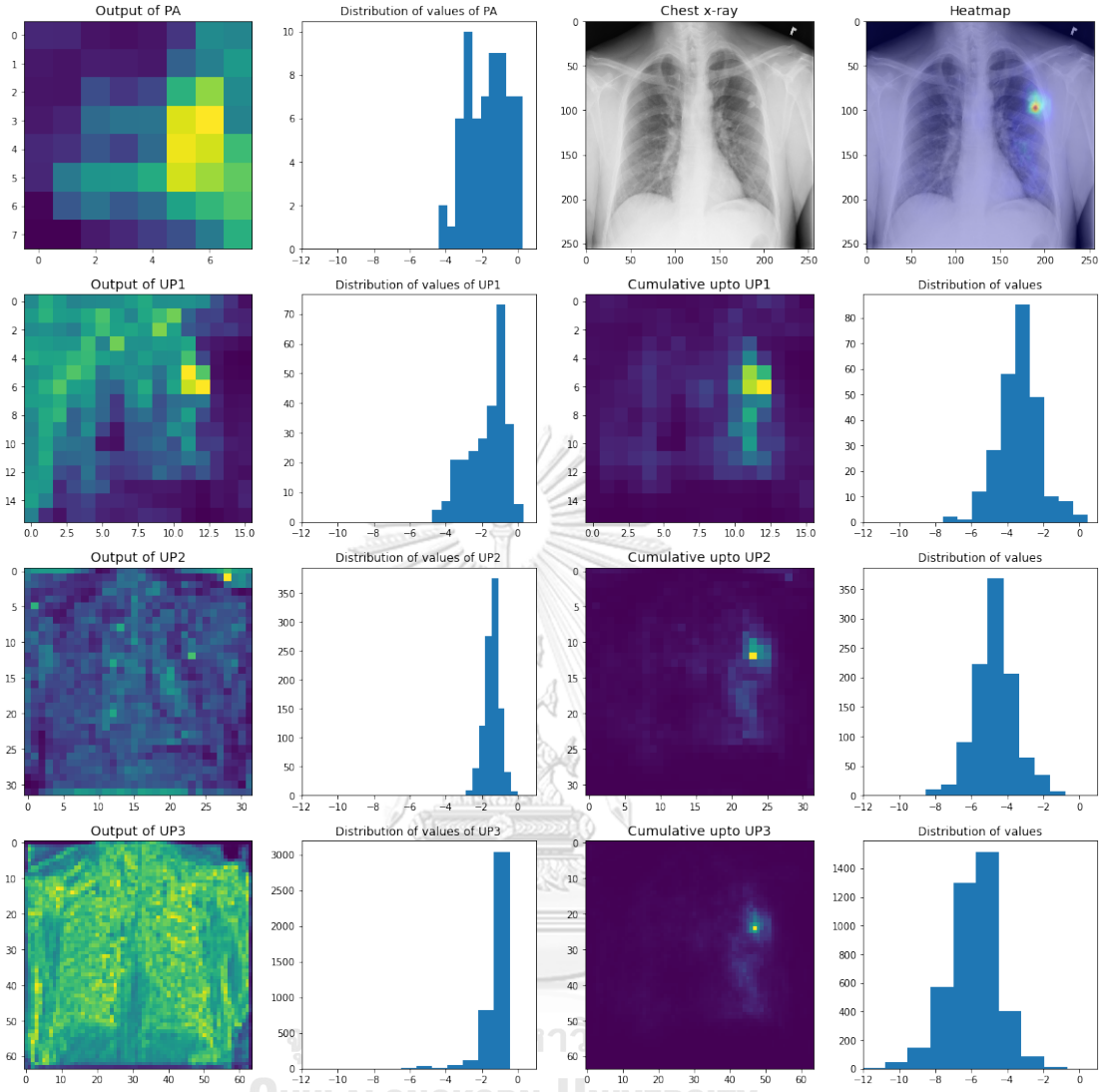


Figure 7.2: A layer-wise perspective on PYLON's heatmap. Showing a heatmap form each component of PYLON and the cumulative heatmap up to each layer. Histograms of values are also presented with the same range. The right-most column histograms show the *negative shifting* of values as the heatmap is added by subsequent layers.

PYLON has *one* PA module and *three* UP modules. We captured outputs from these *four* modules and visualized them in Figure 7.2. The left-most column shows the outputs from all modules accordingly. Outputs of PA and UP1 are reasonably straightforward to understand, they focus on the Nodule site where the output of UP1 has higher resolution. The same cannot be said for the outputs of UP2 and UP3, especially UP3 where the focus areas are spread across the whole x-ray, not specific to the Nodule site. In case of UP2, we can see a *few* hot spots at the Nodule site and other sites not related to nodules in particular. In short, higher layers (PA and UP1) point to the Nodule site with specificity while lower layers (UP2 and UP3) tend to be more spread out focusing multiple sites.

When taking into account the fact that the lower modules (UP2 and UP3) take *lateral* connections from *early* blocks of the encoder which have *lower* capacity and *narrower* receptive field (the upper bound area it can look), we can understand the behavior of the lower UP2 and UP3 better. The lower modules cannot identify the Nodule site by themselves due to the lack of encoder's capacity. This does not mean that UP2 and UP3 are useless because we still see increasing quality of CAM from Figure 7.2 (cumulative). UP2 and UP3 do not work by *pointing* but they work by *negating* areas which are certainly not the Nodule site instead. This is supported by the output of UP2 where there are more than one hotspots, only one of them is the Nodule site, but outside those hotspots are areas quite certain not to be Nodule. The corollary of this finding is that UP2 and UP3 can be made powerful with deeper and higher capacity backbones. Here we used ResNet-50, but one can use DenseNet-169 which is deeper and may yield a more powerful encoder at the lateral connections to UP2 and UP3 for instance.

Another notable observation is on the *negative shifting* of values (the right-most column histograms in Figure 7.2). As the subsequent layer's outputs (with high resolution and lower capacity) are added to the heatmap, the values on the heatmap shifts toward the negative side suggesting that each subsequent layers are mostly *subtracting* values, not adding. This supports the hypothesis we just made above that the UP2 and UP3 work by negating values of unlikely areas of the heatmap. Consequently, PYLON will predict with *lower* confidence than it really should. This does not mean lower prediction accuracy, it means that a threshold should always be properly selected not by a simple value such as 0.5.

7.3 Channel-wise perspective

PYLON constructs heatmaps from linear combinations of multiple *channels*, default is 128, via its prediction head. We take a closer look into each channel before they are combined. Figure 7.3 selectively shows top 20 channels according to the prediction head's *weights* and shows heatmaps with just these 20 channels (from the total of 128) which is not a very good heatmap for the Nodule yet but somewhat resembles the final heatmap. From these 20 channels, we observe *no* channel in particular that demonstrates a clear Nodule characteristic. We will be hard pushed to interpret any channel regarding the Nodule class. Another observation is that the top 20 weights are close in value ranging from 0.17 to 0.23 (we inverted channels with negative weights). This suggests that channels are allocated in a *distributed* way which means it does not rely on any particular channel to produce the heatmap, and only when much of the 128 channels are combined, the heatmap becomes intelligible. Figure 7.4 shows the *cumulative* heatmaps of these 20 channels. We observe that the cumulative heatmap begins with a large focus area around the left lung, yet with increasing number of channels, it becomes increasingly more specific, not by much between steps, but finally reaching the final heatmap nonetheless.

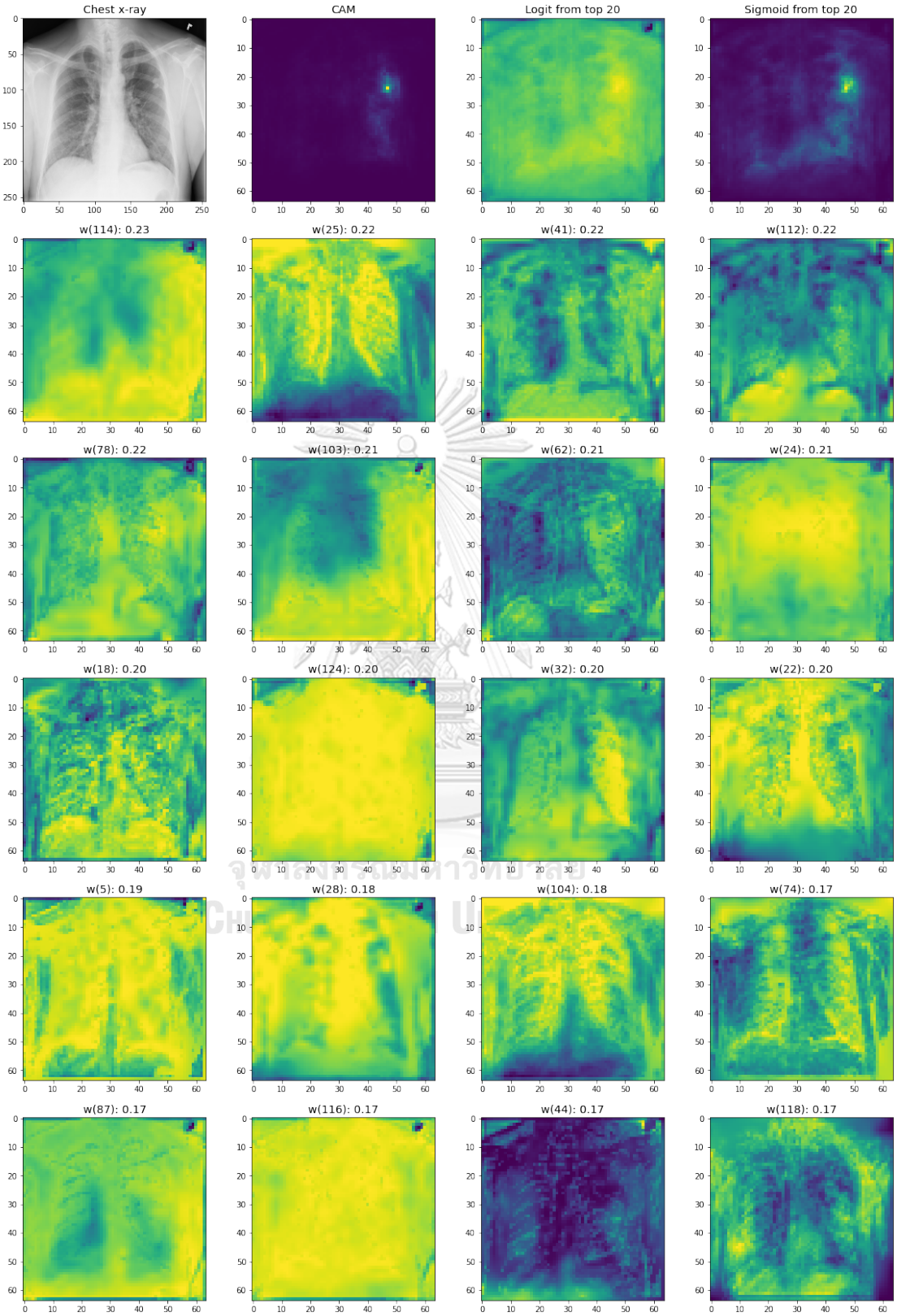


Figure 7.3: A channel-wise perspective on PYLON's heatmap. Showing each channel before they are combined to get the final heatmap. Showing the top 20 channels sorted descendingly by their weights.

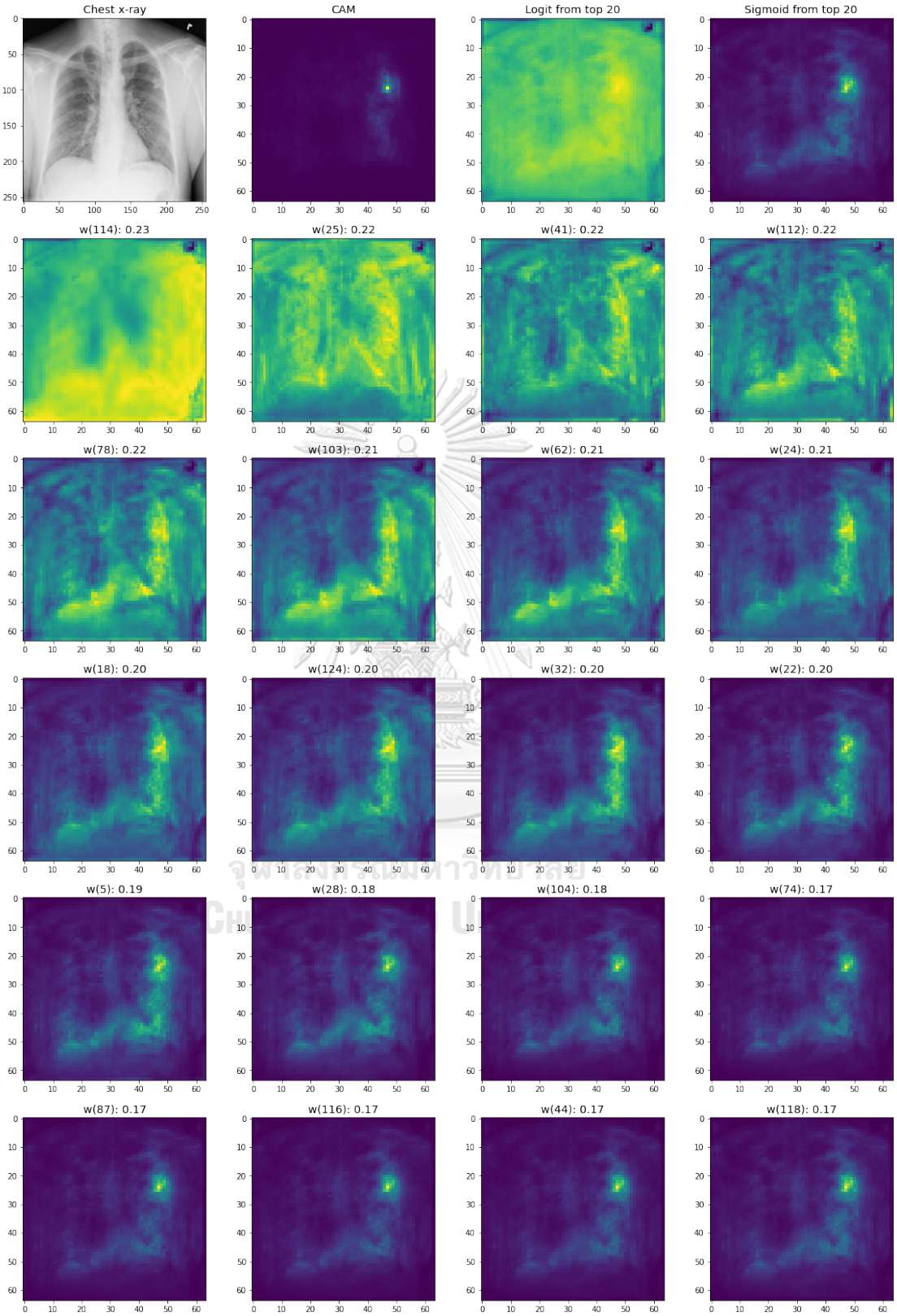


Figure 7.4: Showing the progression of heatmap after taking each decreasingly important channel into account. Showing top 20 channels sorted descendingly by their weights.

Chapter VIII

DISCUSSION

In this study, we proposed PYLON, a deep network architecture for high resolution and high accuracy class activation maps (CAM) to facilitate human interpretation of the model's prediction output. We evaluated PYLON on the publicly available NIH's Chest X-Ray 14 and VinDR-CXR datasets, and compared its performance against previous works and strong segmentation models like PAN (Li et al., 2018a), FPN (Kirillov et al., 2019) and DeeplabV3+ (Chen et al., 2018). Our results indicated that PYLON produces the best heatmaps with respect to the point localization accuracy while maintaining the same level of chest x-ray abnormality classification performance as existing models. PYLON substantially improves CAM accuracy across most abnormality classes especially for small lesions like Nodule (Table 5.1). We also proposed the two-phase transfer learning procedure that further improved the performance of PYLON on small image datasets (Table 5.5).

Not all high resolution CAMs are created equal. All segmentation-based CAM models in this work produced the *same* high resolution heatmaps, yet their results were wildly different (Table 5.1) where PYLON consistently came on top and DeeplabV3+ (Chen et al., 2018) received no advantage from its higher resolution CAM. This suggests that though high resolution heatmaps are important they do not always come with high accuracy. Resolution and accuracy are two different things, yet PYLON seems to deliver both.

Unintelligible CAMs due to group norm. We have found that the original FPN (Kirillov et al., 2019) with *group* normalization produced *unintelligible* heatmaps (Figure 8.1) while having the classification performance on par with the other baselines. We have confirmed the phenomenon with other group sizes and with full precision floating point training procedure. This problem was only alleviated by substituting group normalization with *batch* normalization. We did not expect a normalization layer to have such a negative effect on the CAM's quality when there is no negative effect observable on its classification performance. We cannot yet explain the phenomena about which could be an interesting avenue for further investigation.

Global pooling destroys CAM information. Although global average pooling (GAP) has found to improve overall performance when used as a part of attention mechanisms in both classification networks (Hu et al., 2018) and segmentation networks (Li et al., 2018a; Chen et al., 2018), our study suggests that such mechanisms need to be used with care when the accuracy of CAM is of concern. We have found that GAP, adopted in PAN (Li et al., 2018a) and DeeplabV3+ (Chen et al., 2018), was a destabilizing factor for CAM by showing that their respective models without GAP were more stable and produced better quality heatmaps in general (Table 6.1). A possible explanation follows directly from the main characteristic of *any* global pooling layer

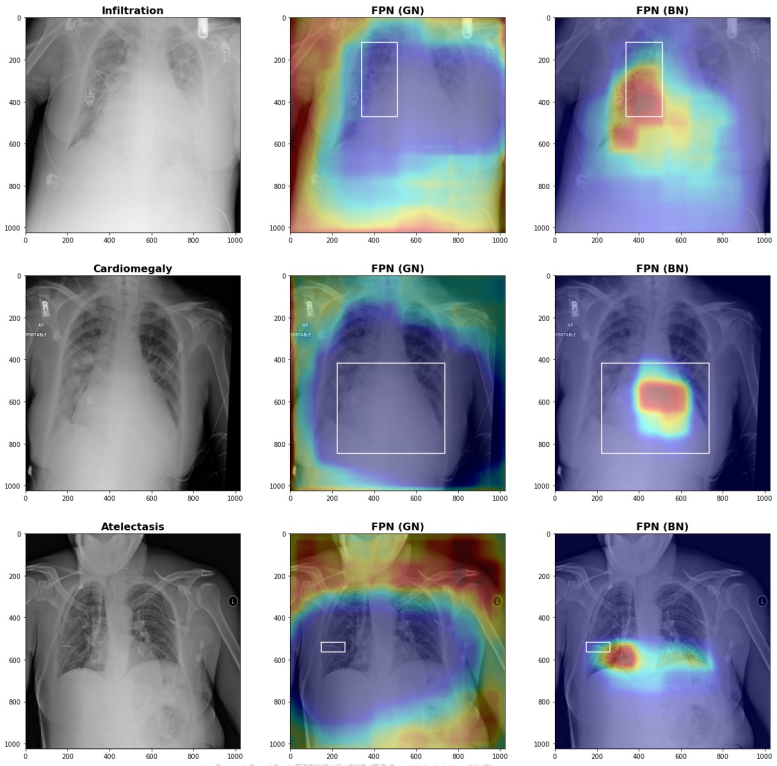


Figure 8.1: FPN with group norm produced unintelligible heatmaps. Images from NIH's Chest X-Ray 14.

which is the collapsing of spatial dimensions. The signal coming out of the global pooling has *no spatial information left* and cannot be recovered to construct accurate CAM in later stages. Hence, we expect our finding to apply beyond GAP to other global pooling choices. However, deep models in practice have multiple paths, through which information is propagated, while *only some* paths may contain global pooling. This potentially leaves some spatial information through. It might explain the high variance between experiments as some experiments happened to have more spatial information left for the CAM to pickup than the others. We conclude that *any* use of GAP in PAN-like decoder lead to undesirable outcomes. This finding is the basis on why PYLON particularly refrained from using any GAP in its decoder.

Conv 1×1 is preferable. Conv 3×3 is usually a common choice for kernel size in convolution for both classification models and semantic segmentation models, however our ablation results suggest that Conv 1×1 produced more accurate heatmaps. Whenever a Conv 3×3 was used instead of a Conv 1×1, there was a drop in CAM accuracy (Table 6.2). CAM relies on the *fidelity* of each layer in the deep network to produce high values at the location corresponding to where it sees. For example, when a model sees a nodule, we want the model to output high values at the precise location of the nodule. Only models with this property would produce intelligible CAM. We argue that Conv 1×1 satisfies this property *stronger* than Conv 3×3. Due to its narrow field of input, its outputs *must* correspond to its input at that precise location. There are limitations to this conclusion. We did not test the hypotheses exhaustively on all known

models. One should refrain from prematurely concluding that this hypothesis applies in general cases with certainty.

Partial discovery of abnormality sites. A closer look at the CAM generated suggests that CAM does not give equal emphasis on all abnormality sites. CAM might focus on a specific site and ignores the rest (Figure 8.2). This is well known in the CAM literature (Huang et al., 2020; Bae et al., 2020) that CAM usually only focuses on the most *discriminative* part of the image. Understandably, a model does not need to glean on all abnormality sites to conclude that the abnormality is present. Without any extra knowledge on how many sites there are in the image, a model trained with only image-level annotation cannot guarantee the discovery of all sites related to a given class.

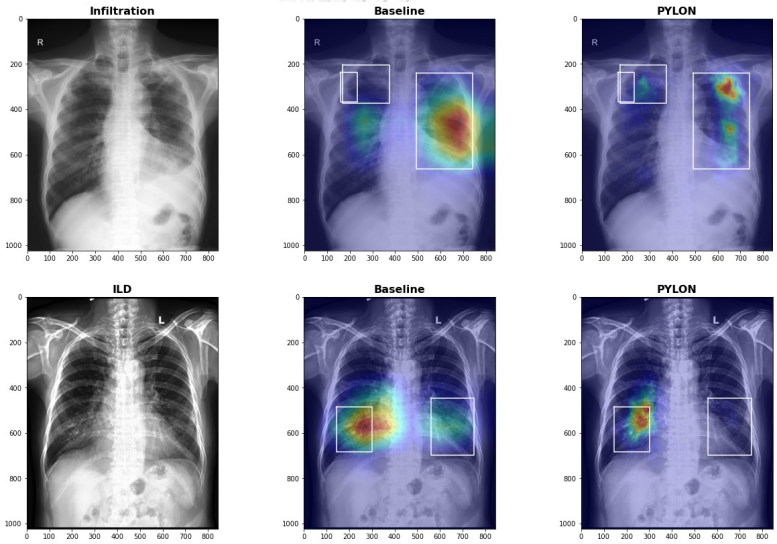


Figure 8.2: CAM cannot guarantee full discovery of all abnormality sites. Images from VinDR-CXR.

Limitations of PYLON. PYLON had been shown to work on the domain of chest radiographs in this work. It is not a far-fetched hypothesis that it should work on a wider range of domain in medical imaging, e.g. CT scan. Since PYLON was not designed with any *medical specific* component, one could potentially use PYLON on non-medical images as well for high-accuracy localization. However, users of PYLON should be aware of its limitations which are CAM's limitations. PYLON (and CAM) is a *partial* explainability method. PYLON does explain the decision made by the encoder perfectly. This can be observed in Cardiomegaly class (Figure 5.1) whose CAM should focus on both the heart and the perimeter of the thorax, not just the heart, in order to determine the cardiothoracic ratio which is the main criterion for Cardiomegaly class. This suggests that CAM does not perfectly reflect how the model looks at the image. PYLON constructs heatmaps that point to the location of *prominent* cues in the input regarding a given class of interest. Fortunately, our chest x-ray classification task is already useful with the explainability of CAM methods by pointing to the region of interests for further investigations by radiologists.

REFERENCES

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. 2018. Sanity checks for saliency maps. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (ed.), Advances in Neural Information Processing Systems 31 (NIPS 2018), pp. 9505–9515. : Curran Associates, Inc.

Ahn, J., Cho, S., and Kwak, S. 2019. Weakly supervised learning of instance segmentation with Inter-Pixel relations. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2204–2213. :

Arora, S., Li, Z., and Lyu, K. 2018. Theoretical analysis of auto Rate-Tuning by batch normalization. (December 2018):

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. 2015. On Pixel-Wise explanations for Non-Linear classifier decisions by Layer-Wise relevance propagation. PLoS One 10.7 (July 2015): e0130140.

Bae, W., Noh, J., and Kim, G. 2020. Rethinking class activation mapping for weakly supervised object localization. In Computer Vision – ECCV 2020, Lecture notes in computer science, pp. 618–634. Cham: Springer International Publishing.

Bahdanau, D., Cho, K., and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. (September 2014): 1–15.

Baltruschat, I. M., Nickisch, H., Grass, M., Knopp, T., and Saalbach, A. 2019. Comparison of deep learning approaches for Multi-Label chest X-Ray classification. Scientific Reports 9.1 (April 2019): 6381.

Bjorck, N., Gomes, C. P., Selman, B., and Weinberger, K. Q. 2018. Understanding batch normalization. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (ed.), Advances in Neural Information Processing Systems 31, pp. 7694–7705. : Curran Associates, Inc.

Bustos, A., Pertusa, A., Salinas, J.-M., and de la Iglesia-Vayá, M. 2020. PadChest: A large chest x-ray image dataset with multi-label annotated reports. Med. Image Anal. 66 (August 2020): 101797.

Chang, C.-H., Creager, E., Goldenberg, A., and Duvenaud, D. 2018. Explaining image classifiers by counterfactual generation. (July 2018):

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. 2014. Semantic image segmentation with deep convolutional nets and fully connected CRFs. (December 2014):

Chen, L.-C., Papandreou, G., Schroff, F., and Adam, H. 2017a. Rethinking atrous convolution for semantic image segmentation. (June 2017):

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. 2018. Encoder-Decoder with atrous separable convolution for semantic image segmentation. In Computer Vision – ECCV 2018, pp. 833–851. : Springer International Publishing.

Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., and Chua, T.-S. 2017b. SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5659–5667. : IEEE.

Choe, J. and Shim, H. 2019. Attention-Based dropout layer for weakly supervised object localization. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2214–2223. :

Chorowski, J., Bahdanau, D., Serdyuk, D., Cho, K., and Bengio, Y. 2015. Attention-Based models for speech recognition. (June 2015): 1–19.

DeVries, T. and Taylor, G. W. 2017. Improved regularization of convolutional neural networks with cutout. (August 2017):

Fan, L., Zhao, S., and Ermon, S. 2017. Adversarial localization network. In Learning with limited labeled data: weak supervision and beyond, NIPS Workshop. : lijiefan.me.

Fong, R. C. and Vedaldi, A. 2017. Interpretable explanations of black boxes by meaningful perturbation. In 2017 IEEE International Conference on Computer Vision (ICCV), pp. 3449–3457. : IEEE.

Frye, C., Rowat, C., and Feige, I. 2019. Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. (October 2019):

Ghiasi, G., Lin, T.-Y., and Le, Q. V. 2018. DropBlock: A regularization method for convolutional networks. (October 2018):

Guan, Q., Huang, Y., Zhong, Z., Zheng, Z., Zheng, L., and Yang, Y. 2018. Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification. (January 2018):

He, K., Zhang, X., Ren, S., and Sun, J. 2016a. Deep residual learning for image recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. :

He, K., Zhang, X., Ren, S., and Sun, J. 2015. Delving deep into rectifiers: Surpassing Human-Level performance on ImageNet classification. (February 2015):

He, K., Zhang, X., Ren, S., and Sun, J. 2016b. Identity mappings in deep residual networks. In Computer Vision – ECCV 2016, volume 9908 LNCS, pp. 630–645. : Springer International Publishing.

Hu, J., Shen, L., and Sun, G. 2018. Squeeze-and-Excitation networks. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7132–7141. :

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. 2017. Densely connected convolutional networks. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2261–2269. :

Huang, Z., Wang, X., Wang, J., Liu, W., and Wang, J. 2018. Weakly-Supervised semantic segmentation network with deep seeded region growing. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7014–7023. :

Huang, Z., Zou, Y., Bhagavatula, V., and Huang, D. 2020. Comprehensive attention Self-Distillation for Weakly-Supervised object detection. (October 2020):

Ioffe, S. and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the 32nd International Conference on Machine Learning (ICML 2015), ICML'15, pp. 448–456. : JMLR.org.

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D. A., Halabi, S. S., Sandberg, J. K., Jones, R., Larson, D. B., Langlotz, C. P., Patel, B. N., Lungren, M. P., and Ng, A. Y. 2019. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. AAAI 33 (July 2019): 590–597.

Johnson, A. E. W., Pollard, T. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-Y., Peng, Y., Lu, Z., Mark, R. G., Berkowitz, S. J., and Horng, S. 2019. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. (January 2019):

Kingma, D. P. and Ba, J. 2015. Adam: A method for stochastic optimization. In 3rd International Conference on Learning Representations, ICLR 2015. :

Kirillov, A., Girshick, R., He, K., and Dollar, P. 2019. Panoptic feature pyramid networks. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6392–6401. : IEEE.

Krähenbühl, P. and Koltun, V. 2011. Efficient inference in fully connected CRFs with gaussian edge potentials. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F., and Weinberger, K. Q. (ed.), Advances in Neural Information Processing Systems 24 (NIPS 2011), pp. 109–117. : Curran Associates, Inc.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. 2012. ImageNet classification with deep convolutional neural networks. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q. (ed.), Advances in Neural Information Processing Systems 25, pp. 1097–1105. : Curran Associates, Inc.

Li, H., Xiong, P., An, J., and Wang, L. 2018a. Pyramid attention network for semantic segmentation (BMVC 2018). In 29th British Machine Vision Conference (BMVC 2018). :

Li, K., Wu, Z., Peng, K.-C., Ernst, J., and Fu, Y. 2018b. Tell me where to look: Guided attention inference network. (February 2018):

Li, L., Qin, L., Xu, Z., Yin, Y., Wang, X., Kong, B., Bai, J., Lu, Y., Fang, Z., Song, Q., Cao, K., Liu, D., Wang, G., Xu, Q., Fang, X., Zhang, S., Xia, J., and Xia, J. 2020. Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT. Radiology (March 2020): 200905.

Li, Z., Wang, C., Han, M., Xue, Y., Wei, W., Li, L., and Fei-Fei, L. 2018c. Thoracic disease identification and localization with limited supervision. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8290–8299. :

Lin, T., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. 2017. Feature pyramid networks for object detection. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), volume 2017-Janua, pp. 936–944. :

Lin, Z. Q., Shafiee, M. J., Bochkarev, S., St. Jules, M., Wang, X. Y., and Wong, A. 2019. Do explanations reflect decisions? a machine-centric strategy to quantify the performance of explainability algorithms. (October 2019):

Liu, G., Hsu, T.-M. H., McDermott, M., Boag, W., Weng, W.-H., Szolovits, P., and Ghassemi, M. 2019a. Clinically accurate chest X-Ray report generation. In Doshi-Velez, F., Fackler, J., Jung, K., Kale, D., Ranganath, R., Wallace, B., and Wiens, J. (ed.), Proceedings of the 4th Machine Learning for Healthcare Conference, PMLR, volume 106 of Proceedings of Machine Learning Research, pp. 249–269. Ann Arbor, Michigan: PMLR.

Liu, J., Zhao, G., Fei, Y., Zhang, M., Wang, Y., and Yu, Y. 2019b. Align, attend and locate: Chest X-Ray diagnosis via contrast induced attention network with limited supervision. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10631–10640. :

Long, J., Shelhamer, E., and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3431–3440. :

Lundberg, S. M. and Lee, S.-I. 2017. A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (ed.), Advances in Neural Information Processing Systems 30 (NIPS 2017), pp. 4765–4774. : Curran Associates, Inc.

Nguyen, H. Q., Lam, K., Le, L. T., Pham, H. H., Tran, D. Q., Nguyen, D. B., Le, D. D., Pham, C. M., Tong, H. T. T., Dinh, D. H., Do, C. D., Doan, L. T., Nguyen, C. N., Nguyen, B. T., Nguyen, Q. V., Hoang, A. D., Phan, H. N., Nguyen, A. T., Ho, P. H., Ngo, D. T., Nguyen, N. T., Nguyen, N. T., Dao, M., and Vu, V. 2020. VinDr-CXR: An open dataset of chest x-rays with radiologist's annotations. (December 2020):

Oquab, M., Bottou, L., Laptev, I., and Sivic, J. 2015. Is object localization for free? - weakly-supervised learning with convolutional neural networks. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 685–694. :

Peng, Y., Wang, X., Lu, L., Bagheri, M., Summers, R., and Lu, Z. 2018. NegBio: a high-performance tool for negation and uncertainty detection in radiology reports. AMIA Jt Summits Transl Sci Proc 2017 (May 2018): 188–196.

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M. P., and Ng, A. Y. 2017. CheXNet: Radiologist-Level pneumonia detection on chest X-Rays with deep learning. (November 2017): 3–9.

Ren, Z., Yu, Z., Yang, X., Liu, M.-Y., Lee, Y. J., Schwing, A. G., and Kautz, J. 2020. Instance-aware, context-focused, and memory-efficient weakly supervised object detection. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). : IEEE.

Ronneberger, O., Fischer, P., and Brox, T. 2015. U-Net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, pp. 234–241. : Springer International Publishing.

Rozenberg, E., Freedman, D., and Bronstein, A. 2020. Localization with limited annotation for chest x-rays. In Dalca, A. V., McDermott, M. B. A., Alsentzer, E., Finlayson, S. G., Oberst, M., Falck, F., and Beaulieu-Jones, B. (ed.), Proceedings of the Machine Learning for Health NeurIPS Workshop 2019, volume 116 of Proceedings of Machine Learning Research, pp. 52–65. : PMLR.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. 2015. ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. 115.3 (December 2015): 211–252.

Santurkar, S., Tsipras, D., Ilyas, A., and Madry, A. 2018. How does batch normalization help optimization?. (May 2018):

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. 2017. Grad-CAM: Visual explanations from deep networks via Gradient-Based localization. In 2017 IEEE International Conference on Computer Vision (ICCV), pp. 618–626. :

Shin, H., Roberts, K., Lu, L., Demner-Fushman, D., Yao, J., and Summers, R. M. 2016. Learning to read chest X-Rays: Recurrent neural cascade model for automated image annotation. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2497–2506. :

Shrikumar, A., Greenside, P., Shcherbina, A., and Kundaje, A. 2016. Not just a black box: Learning important features through propagating activation differences. (May 2016):

Shrikumar, A., Greenside, P., and Kundaje, A. 2017. Learning important features through propagating activation differences. In Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17, pp. 3145–3153. : JMLR.org.

Simonyan, K. and Zisserman, A. 2015. Very deep convolutional networks for Large-Scale image recognition. In Bengio, Y. and LeCun, Y. (ed.), 3rd International Conference on Learning Representations, ICLR 2015. :

Simonyan, K., Vedaldi, A., and Zisserman, A. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013): 1–8.

Singh, K. K. and Lee, Y. J. 2017. Hide-and-Seek: Forcing a network to be meticulous for Weakly-Supervised object and action localization. In 2017 IEEE International Conference on Computer Vision (ICCV), pp. 3544–3553. :

Su, H., Jampani, V., Sun, D., Gallo, O., Learned-Miller, E., and Kautz, J. 2019. Pixel-Adaptive convolutional neural networks. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11158–11167. :

Tan, M. and Le, Q. 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. In Chaudhuri, K. and Salakhutdinov, R. (ed.), Proceedings of the 36th International Conference on Machine Learning, PMLR, volume 97 of Proceedings of Machine Learning Research, pp. 6105–6114. Long Beach, California, USA: PMLR.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., and Summers, R. M. 2017. 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. :

Wang, X., Peng, Y., Lu, L., Lu, Z., and Summers, R. M. 2018. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9049–9058. :

Wu, Y. and He, K. 2018. Group normalization. In Computer Vision – ECCV 2018, pp. 3–19. : Springer International Publishing.

Yao, L., Poblenz, E., Dagunts, D., Covington, B., Bernard, D., and Lyman, K. 2017. Learning to diagnose from scratch by exploiting dependencies among labels. (October 2017):

Yao, L., Prosky, J., Poblenz, E., Covington, B., and Lyman, K. 2018. Weakly supervised medical diagnosis and localization from multiple resolutions. (March 2018):

Zeiler, M. D. and Fergus, R. 2014. Visualizing and understanding convolutional networks. In Computer Vision – ECCV 2014, pp. 818–833. : Springer International Publishing.

Zhang, C., Chen, F., and Chen, Y.-Y. 2020. Thoracic disease identification and localization using distance learning and region verification. (June 2020):

Zhang, J., Bargal, S. A., Lin, Z., Brandt, J., Shen, X., and Sclaroff, S. 2018a. Top-Down neural attention by excitation backprop. Int. J. Comput. Vis. 126.10 (October 2018): 1084–1102.

Zhang, R. 2019. Making convolutional networks Shift-Invariant again. In Proceedings of the 36th International Conference on Machine Learning, PMLR. :

Zhang, X., Wei, Y., Feng, J., Yang, Y., and Huang, T. 2018b. Adversarial complementary learning for weakly supervised object localization. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1325–1334. :

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. 2016. Learning deep features for discriminative localization. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2921–2929. :

Zhu, Y., Zhou, Y., Ye, Q., Qiu, Q., and Jiao, J. 2017. Soft proposal networks for weakly supervised object localization. In 2017 IEEE International Conference on Computer Vision (ICCV), pp. 1859–1868. :

Zolna, K., Geras, K. J., and Cho, K. 2020. Classifier-agnostic saliency map extraction. Computer Vision and Image Understanding 196 (July 2020): 102969.

The person on the left is me, **Konpat Preechakul (Ta)**.
I was a computer engineer turned machine learning researcher.
Computer has always been my love since I was a small child who picked up my father's "คู่มือประกอบคอมพิวเตอร์" book left on a table. I have read it cover-to-cover and was fascinated by it. Since then, it was clear to me to pursue goals in the field of computer. I was fortunate enough to get into a computer olympiad programme. It also gave me a ticket to attend to Chulalongkorn University which I was not capable of attending it otherwise.
Only at the later years of my bachelor study would I have cared about the idea of *intelligent* machines. I was hooked by the idea of *prosperity without work* which, I am convinced, only realizable via the path to intelligent machines. Ever since, my life-long goal has been to *rid* all people of jobs. Though I do not agree with his method, I am enchanted by his vision that some day in the future:
"it possible for me to do one thing today and another tomorrow, to hunt in the morning, fish in the afternoon, rear cattle in the evening, criticise after dinner, just as I have a mind, without ever becoming hunter, fisherman, herdsman or critic."
— Karl Marx