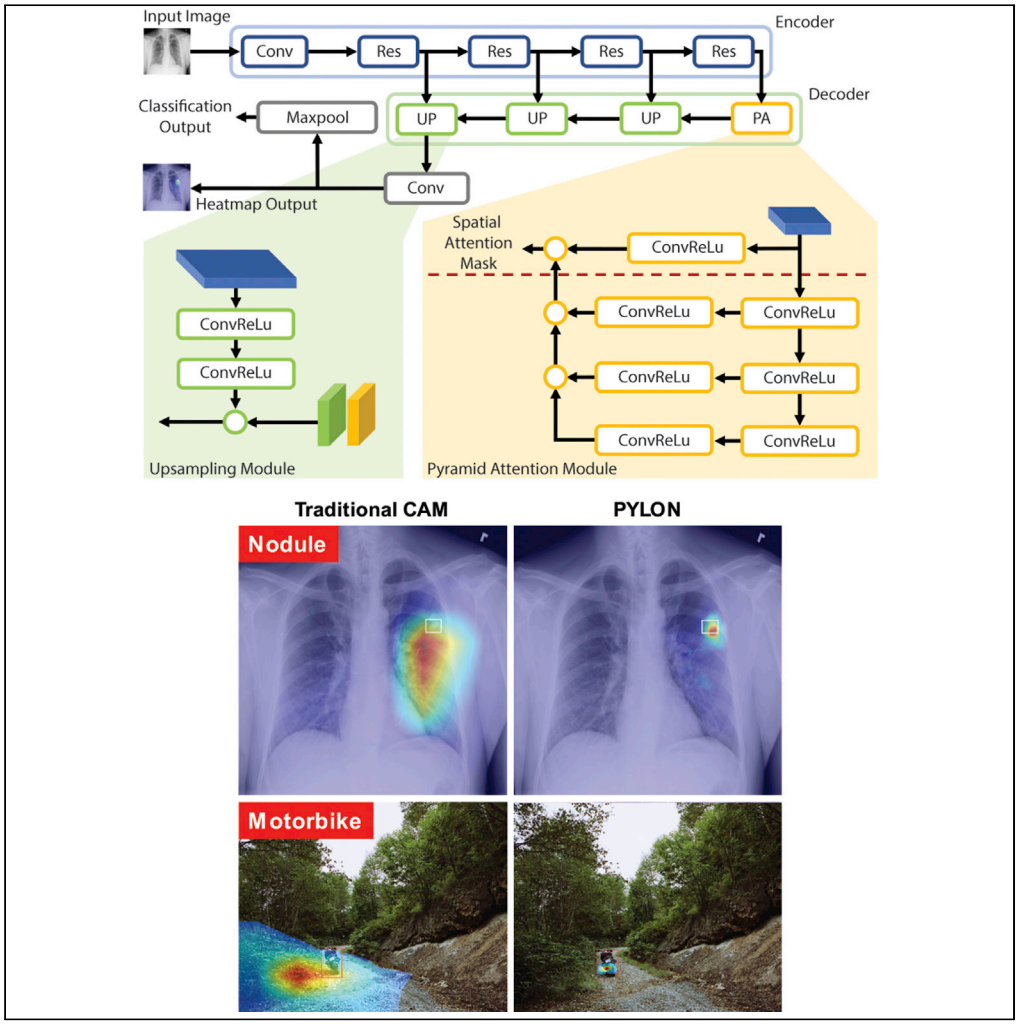
iScience

CellPress
OPEN ACCESS

Article

Improved image classification explainability with high-accuracy heatmaps



Konpat Preechakul, Sira Sriswasdi, Boonserm Kijsirikul, Ekapol Chuangsuwanich

konpatp@gmail.com (K.P.)
sira.sr@chula.ac.th (S.S.)
ekapol.c@chula.ac.th (E.C.)

Highlights

New architecture improves accuracy of heatmaps that explain classification model

Accurately pinpoint small objects, such as nodules in chest radiograph

No pixel-level annotation is required for training

Applicable on small dataset with a two-phase transfer learning approach

Preechakul et al., iScience 25, 103933
March 18, 2022
https://doi.org/10.1016/j.isci.2022.103933

iScience

CellPress
OPEN ACCESS

Article

Improved image classification explainability with high-accuracy heatmaps

Konpat Preechakul,[1,2,*] Sira Sriswasdi,[2,3,4,*] Boonserm Kijsirikul,[1] and Ekapol Chuangsuwanich[1,2,5,*]

SUMMARY

Deep learning models have become increasingly used for image-based classification. In critical applications such as medical imaging, it is important to convey the reasoning behind the models' decisions in human-understandable forms. In this work, we propose Pyramid Localization Network (PYLON), a deep learning model that delivers precise location explanation by increasing the resolution of heatmaps produced by class activation map (CAM). PYLON substantially improves the quality of CAM's heatmaps in both general image and medical image domains and excels at pinpointing the locations of small objects. Most importantly, PYLON does not require expert annotation of the object location but instead can be trained using only image-level label. This capability is especially important for domain where expert annotation is often unavailable or costly to obtain. We also demonstrate an effective transfer learning approach for applying PYLON on small datasets and summarize technical guidelines that would facilitate wider adoption of the technique.

INTRODUCTION

Deep learning models exhibited strong performances in general image classification (He et al., 2015; Russakovsky et al., 2015) and have become widely used in medical imaging especially for automated diagnoses of chest radiographs (Rajpurkar et al., 2017; Irvin et al., 2019; Sim et al., 2020). However, while deep models provide high classification accuracy, their complexities came at the cost of explainability. Although many techniques have been developed to help humans understand deep models (Simonyan et al., 2013; Bach et al., 2015; Shrikumar et al., 2017; Ancona et al., 2017; Zhang et al., 2018a; Mopuri, Garg, and Venkatesh Babu 2019), they rely on either analysis of gradient, which tends to be difficult to interpret, or probing deep into the models, which is highly specific to the model architecture, or both. Nowadays, especially in the medical domain, the explainability, or interpretability, of deep learning models has become almost as important as their performance. This is because the high complexity of deep learning models makes it prone to spurious associations and biases, allowing the models to achieve high performance by exploiting features unrelated to the task objectives. If left unchecked, such models will make high confidence but erroneous predictions that may have costly consequences.

One common way to explain a deep model's prediction given an input image is through a heatmap where a color highlights important regions on the image that corresponds to a region of interest. Class activation map (CAM) is a popular family of techniques (Oquab et al., 2015; Zhou et al., 2016; Selvaraju et al., 2017; Chattopadhyay et al., 2017; Wang et al., 2019; Muhammad and Yeasin, 2020; Desai and Ramaswamy 2020; Fu et al., 2020) which derives class-specific heatmaps as by-products of the prediction process inside the models. These techniques rely on the multi-instance learning (MIL) principles (Dietterich et al., 1997; Cinbis et al., 2017) which state that each decision made by the model must be attributable to at least one evidence in the input image. By design, CAM focuses on explaining the top-most layers of a deep model which are relatively easier to interpret for human users. CAM methods were first invented for explaining deep convolutional models in the general image domains. Pioneered by Wang et al. (Wang et al., 2017) and Rajpurkar et al. (Rajpurkar et al., 2017) to use CAM-generated heatmaps to point out likely pathological regions in chest radiographs, CAM has become a common approach in the medical image domain.

High-resolution heatmaps are important because they can clearly convey whether the model utilized the correct pixels from the object to derive its classification output, as supposed to nearby unrelated pixels.

[1]Department of Computer Engineering, Chulalongkorn University, Pathum Wan, Bangkok 10330, Thailand

[2]Center of Excellence in Computational Molecular Biology, Faculty of Medicine, Chulalongkorn University, 254 Phayathai Road, Pathum Wan, Bangkok 10330, Thailand

[3]Research Affairs, Faculty of Medicine, Chulalongkorn University, Pathum Wan, Bangkok 10330, Thailand

[4]Center for Artificial Intelligence in Medicine, Faculty of Medicine, Chulalongkorn University, Pathum Wan, Bangkok 10330, Thailand

[5]Lead contact

*Correspondence: konpatp@gmail.com (K.P.), sira.sr@chula.ac.th (S.S.), ekapol.c@chula.ac.th (E.C.)

https://doi.org/10.1016/j.isci.2022.103933

With high-resolution, precise heatmaps, the human operators can easily determine whether the prediction is trustworthy by verifying that the objects of interest are present in the heatmap regions. Hence, the more precise the heatmaps are, the easier it is to notice when a prediction error is made. However, existing CAM methods produce low-resolution heatmaps. This is because CAM heatmaps are produced from the highest-level feature maps of deep classifiers that are much smaller than the input image size. In most cases, these low-resolution heatmaps are simply upscaled with no improvement in resolution to be overlaid on top of the input image as a form of explanation. Although there are many improvements in CAM methods over the years (Chattopadhyay et al., 2017; Wang et al., 2019; Muhammad and Yeasin, 2020; Desai and Ramaswamy 2020; Fu et al., 2020), none addresses the problem of a low-resolution heatmap. For chest radiograph classification, the low resolution 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. This limitation also applies to general image domains where the low-resolution heatmap hinders the ability to localize small instances of an object class. Although one can arbitrarily increase the resolution of CAM heatmaps by increasing the input image size, this will also greatly increase the computational costs.

Here, we propose Pyramid Localization Network (PYLON) that effectively improves the resolution of heatmaps produced by CAM methods without requiring a larger input image size. PYLON accomplishes this by aggregating spatial information from multiple layers inside a deep classifier to produce final high-resolution heatmaps. Extensive evaluations indicated that PYLON considerably improves the accuracy of CAM heatmaps compared to existing methods in both general image and medical image domains. Most importantly, PYLON does not require expert annotation of the object location but instead can be trained using only image-level label. This capability is especially important for medical image domain where expert annotation is often unavailable or costly to obtain. We also demonstrate an effective transfer learning procedure for applying PYLON to small datasets and present a guideline for designing a deep learning model capable of generating accurate heatmaps.

RESULTS

High-resolution heatmaps with Pyramid Localization Network

Class activation maps (CAM) are generated from the deepest convolutional layer of an artificial neural network model which produces low-resolution heatmaps that are often smaller than 1/1000 of the origin input image size. These low-resolution CAM heatmaps are then directly upscaled and overlaid on top of the original input image to serve as the model's explanations. This creates an illusion that the explanations have the same level of resolution as the original image, while, in fact, these broad heatmaps cannot pinpoint the location of the object being classified (Figure 1A, Traditional CAM). In this work, we proposed a Pyramid Localization Network architecture that can generate high-resolution, precise heatmaps (Figure 1A, PYLON). PYLON achieves the high-resolution explanations via a learnable upscaling process that is composed of a Pyramid Attention (PA) and Upsampling (UP) blocks (Figure 1B). This process utilizes the high-resolution but information-scarce signals of the shallower layers of the model to refine the low-resolution but information-rich heatmap from the deeper layers of the model. Most importantly, PYLON requires minimal changes to the existing architectures and does not require object-level annotation to train.

High-resolution heatmaps can convince human operators that the classification output is trustworthy because they clearly convey whether the prediction was influenced by pixels on the actual object or nearby unrelated pixels. Hence, the desirable quality of CAM heatmap in this case is its ability to direct the human operator's attention to the precise object locations. This led us to propose a new metric named point localization accuracy (see STAR Methods), which indicates whether the heatmap pixel with the highest value lies on top of the object, to be evaluated alongside intersection-over-union and masked-out metrics which measure other qualities of the heatmaps. It should be noted that point localization accuracy corresponds to the use of arrows to point out object locations, which is a common practice in chest radiograph reading. Furthermore, the need for high-resolution CAM heatmaps also led us to consider segmentation networks, namely Unet (Ronneberger et al., 2015), FPN (Kirillov et al., 2019), PAN (Li et al., 2018a), and DeeplabV3+ (Chen et al., 2018), that were inherently developed to produce high-resolution outputs, as strong benchmarks for comparing against PYLON.

PYLON performance on NIH's chest X-ray 14

In the medical image domain, NIH's Chest X-ray14 (Wang et al., 2017) has been the main dataset for evaluating CAM accuracy (Wang et al., 2017; Li et al., 2018b; Yao et al., 2018; Rozenberg et al., 2020; Liu et al., 2019) because it contains more than 100,000 images, about 1,000 of which were also annotated with

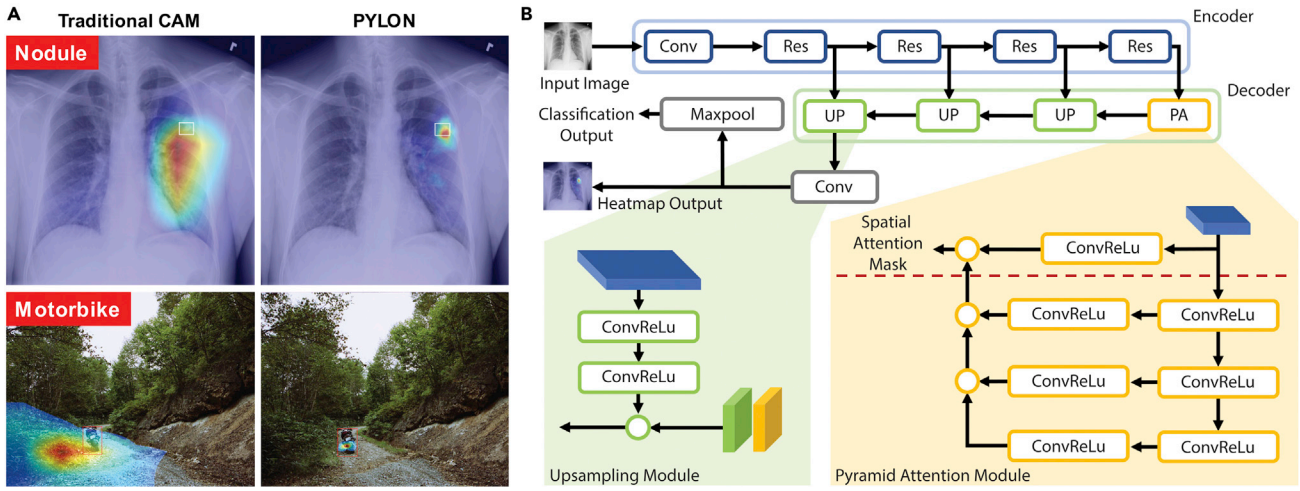


Figure 1. High-resolution heatmap with Pyramid Localization Network (PYLON)

(A) PYLON produces heatmaps with much higher resolutions than those of traditional CAM (Oquab et al., 2015).

(B) The architecture of Pyramid Localization Network (PYLON) with its Pyramid Attention (PA) and Upsampling (UP) modules. PYLON consists of three parts: an encoder, a decoder, and a prediction head. The encoder may be replaced by any deep classifier, such as ResNet or DenseNet. Here, the model variant with an input of size 256 × 256 and ResNet-50 as the encoder was shown. Heatmap is the CAM output. Global Maxpool summarizes class-specific heatmaps into classification outputs. 2X indicates bilinear upsampling. 0.5x indicates 2 × 2 max pooling. Each ConvReLU is a convolution layer followed by batch normalization and ReLU activation. The numbers along with the arrows denote the number of channels while the numbers in parentheses denote the sizes of the feature maps. In the PA module, there is a pyramidal attention path that produces a single-channel spatial attention mask that will be multiplied with the main Conv 1 × 1 path.

bounding boxes to show the locations of abnormalities. The authors also provided a train/test split of this dataset which was also used here to ensure a fair evaluation.

Overall, the heatmaps produced by PYLON can pinpoint the locations of abnormalities in chest radiograph images most accurately among competing techniques (Table 1, 95% CIs are provided in Table S1). The largest performance gaps were found on classes with small object instances, such as 0.5 versus 0.38 point localization accuracy for Atelectasis (average 3% of image area) and 0.48 versus 0.15 point localization accuracy for Nodule (average 1% of image area). The same results held when the input image size was increased from 256 × 256 to 512 × 512, indicating that PYLON consistently improves heatmap qualities (Tables 2 and S2). PYLON's heatmaps also prioritize the most impactful image regions, as captured by the masked-out area metric (see STAR Methods) on 10 of 14 classes (Tables S3 and S4). However, due to the high variance of this metric, the differences are not statistically significant.

In terms of explainability, PYLON's heatmaps are more precise and tend to be confined within the bounding boxes annotated by experts while other models' heatmaps tend to also highlight large area outside the bounding boxes (Figure 2, S1–S4). This explains why the intersection-over-union metric (IoU), which measures the extent of overlap between proposed heatmaps and bounding box annotations, is similarly low across all models (Tables S5 and S6). However, PYLON's low IoU was caused by precise heatmaps that only partially cover the bounding boxes, while other models' low IoUs were due to large heatmaps that spread beyond the annotated regions. This suggests that IoU, which was a standard metric for object detection, is not appropriate for assessing explainability, and that PYLON's heatmsaps are more effective at directing the human operators' attention to the object's true locations.

It should be noted that the classification accuracies of all models are similar on both 256 × 256 and 512 × 512 input sizes (Table S7). An exception was Li 2018 (Li et al., 2018b) that used a different pooling function which could be responsible for the slightly superior classification performance.

Adapting PYLON to small VinDr-CXR datasets with transfer learning

VinDr-CXR (Nguyen et al., 2020) is a relatively new chest radiograph dataset with 18,000 images. As the official test dataset was not publicly available, we split the official train dataset of 15,000 images into our train,

Table 1. Point localization accuracy on the NIH's Chest X-ray 14 dataset with input image size of 256x256

Class	Avg.Area	Baseline	GradCAM[a]	GradCAM++[a]	XGradCAM[a]	PAN	Li et al. (2018b)	UNET	FPN	DeeplabV3+	PYLON
Weight. avg.	0.08	0.46	0.46	0.31	0.23	0.38	0.5	0.45	0.53	0.45	0.61
Macro avg.	0.07	0.44	0.44	0.29	0.22	0.37	0.47	0.44	0.5	0.43	0.6
Atelectasis	0.03	0.31	0.3	0.21	0.14	0.24	0.36	0.24	0.39	0.26	0.5
Cardiomegaly	0.18	0.99	0.99	0.57	0.5	0.51	0.96	0.77	1	0.81	0.99
Effusion	0.07	0.32	0.32	0.25	0.24	0.63	0.5	0.39	0.42	0.49	0.54
Infiltration	0.1	0.6	0.6	0.52	0.4	0.35	0.6	0.59	0.63	0.57	0.71
Mass	0.04	0.42	0.42	0.28	0.27	0.63	0.51	0.6	0.59	0.5	0.67
Nodule	0.01	0.09	0.09	0.07	0.07	0.1	0.06	0.15	0.14	0.07	0.48
Pneumonia	0.09	0.61	0.61	0.33	0.02	0.3	0.55	0.61	0.71	0.53	0.71
Pneumothorax	0.05	0.16	0.16	0.11	0.12	0.18	0.18	0.16	0.14	0.22	0.2

CIs are provided in supplemental information.
[a]Variation of GradCAM methods was performed on the Baseline model.

validation, and test sets with ratio of 70:10:20, respectively. It should be noted that images from the same patient may exist across the splits because the patient IDs were not available at the time of this study. Each image was labeled by three different radiologists and the bounding boxes annotated for the same class in the same image were merged into a single ground truth region.

Similar to the evaluation on the NIH dataset, PYLON achieved considerably higher point localization accuracy than others (Tables 3 and S8, 0.42 versus 0.36 macro average, and the highest for 10 of 14 classes). Again, the IoU metric cannot distinguish the quality of heatmaps produced by the Baseline, FPN (batch norm), and PYLON (Tables S9, 0.13–0.15 macro average) even though the heatmaps were clearly qualitatively different (Figures 3, S5, and S6). Compared to the performances on the NIH dataset, all models exhibited much lower heatmap accuracies and qualities likely due to the limited size of the VinDr-CXR dataset. To overcome this issue, we developed a two-phase fine-tuning technique that substantially improved PYLON's macro average point localization accuracy from 0.42 without pre-training and 0.49 with standard transfer learning technique to 0.56 (Table 3). Masked-out area and IoU metrics were also improved by the two-phase technique over standard transfer learning (Tables S9 and S10). It should be noted that classification performance did not change much by the fine-tuning process, likely because the models already achieved very high accuracies (Table S11, AUROC >0.95). In terms of explainability, our fine-tuning technique clearly improved the precision of the heatmaps (Figures 4, S7, and S8), both reducing their spread and increasing their overlap with the ground truth annotations.

PYLON performance on general image domain

To showcase PYLON's impact on the general image domain, we selected the Pascal VOC2012 dataset (Everingham et al., 2011) which had been widely used to benchmark object detection and weakly supervised models (Ahn and Kwak, 2018; Liu et al., 2020; Huang et al., 2020). The train and validation sets contain 5,717 and 5,823 images, respectively. All models were initialized with ImageNet-pretrained weights, and our two-phase fine-tuning technique was applied to PYLON since this dataset is relatively small.

Overall, PYLON with two-phase fine-tuning achieved the best point localization accuracy in 17 out of 20 classes (Tables 4 and S12) and the best masked-out area in 13 out of 20 classes (Table S13). Again, the largest margins of heatmap quality improvement were observed on classes with small objects such as bottle, TV monitor, and potted plants. Interestingly, the IoU and classification accuracy metrics favored the Baseline model on this dataset while advanced CAM methods, GradCAM, GradCAM++, and XGradCAM, did not produce more accurate heatmaps than the Baseline (Tables 4, S14, and S15). This may be because these CAM methods were not specifically designed for binary classification task.

Nonetheless, the higher IoU and classification accuracy of the Baseline model did not translate into better quality heatmaps. It is evident that the Baseline model produced low-resolution heatmaps which are slightly off from the objects' locations and that none of the advanced CAM methods resolved this limitation

Table 2. Point localization accuracy on the NIH's Chest X-ray 14 dataset with input image size of 512x512

Class	Avg.Area	Baseline	GradCAM[a]	GradCAM++[a]	XGradCAM[a]	Li et al. (2018b)	FPN	PYLON
Weight. avg.	0.08	0.58	0.58	0.06	0.29	0.54	0.61	0.65
Macro avg.	0.07	0.55	0.55	0.05	0.28	0.52	0.59	0.63
Atelectasis	0.03	0.47	0.47	0.02	0.29	0.37	0.5	0.55
Cardiomegaly	0.18	0.96	0.96	0.19	0.39	0.95	0.97	0.97
Effusion	0.07	0.58	0.58	0.05	0.48	0.48	0.61	0.61
Infiltration	0.1	0.71	0.71	0.07	0.37	0.65	0.67	0.71
Mass	0.04	0.58	0.58	0.04	0.39	0.52	0.66	0.74
Nodule	0.01	0.17	0.17	0	0.13	0.32	0.36	0.54
Pneumonia	0.09	0.72	0.72	0.06	0.06	0.56	0.74	0.77
Pneumothorax	0.05	0.18	0.18	0.01	0.13	0.31	0.18	0.19

CIs are provided in supplemental information.
[a]Variation of GradCAM methods was performed on the Baseline model.

(Figure 5 and S9–S12). In contrast, PYLON's heatmaps precisely overlap with the object in many cases, making it easy to visually verify that the model made the correct classifications by locating the right objects.

Contributions of PYLON's architecture on heatmap accuracy

PYLON consists of several architecture components designed to improve the resolution and accuracy of CAM heatmaps (Figure 1B). To simplify the analyses, a smaller variant of PYLON with a single Conv 1 × 1 in its UP module was used instead of the proposed PYLON with two Conv 1 × 1 (Figure S14). Ablation analyses were performed only on the NIH's Chest X-ray 14 which is the largest dataset. As expected, the strongest impact on heatmap quality came from the Upsampling (UP) module, with average point localization accuracy improvement from 0.37 without the UP module to 0.57, 0.61, and 0.62 with one, two, or three UP modules, respectively. The choice of a single-layer Conv 1 × 1 without activation in the prediction head also contributed a 0.05 improvement in point localization accuracy over a larger Conv 3 × 3 layer (0.62 versus 0.57) while the classification accuracy only slightly decreased (0.817 versus 0.819). The Pyramid Attention (PA) module additionally improved point localization accuracy by 0.02 (0.60 without PA versus 0.62 with PA). Full results are provided in Table S16.

Overall, changing a Conv 1 × 1 layer to a larger Conv 3 × 3 layer in either the UP module or the prediction head caused a sizable drop in point localization accuracy. This suggests that Conv 1 × 1 might be preferable for generating high-quality CAM in general, possibly because Conv 1 × 1 has a narrower field of view than a Conv 3 × 3 and therefore can better preserve spatial information. Furthermore, to investigate whether the use of heatmap output in conjunction with a simple global maxpooling for the prediction head had any adverse impact on PYLON's classification performance, the prediction head was replaced by a conventional average pooling layer followed by up to three fully connected layers. This shows that PYLON's classification performance did not suffer from the use of the simple global maxpooling output (Tables S17, 0.819 versus 0.820 macro average AUROC). It also suggests that the backbone, not the classification head, was responsible for most of the classification performance.

Adverse impacts of global pooling and group normalization on explainability

Experiments on NIH's Chest X-ray 14 dataset showed that the quality of heatmaps produced by segmentation networks like PAN and DeeplabV3+ exhibited large variations even under the same hyperparameter setting (i.e., the fluctuations were caused by different random initializations). This behavior did not occur in FPN (batch norm) which does not contain a global average pooling (GAP) layer in its architecture. Furthermore, immediate improvement in heatmap quality metrics was achieved for these models by removing GAP from their architecture. Macro average point localization accuracy for PAN improved from 0.37 with SD of 0.16–0.59 to SD of 0.03 when all GAP layers were removed. Removal of GAP layers from DeeplabV3+ also reduced the variance of point localization accuracy in several classes. Adding GAP layers to PYLON also worsen its point localization accuracy by up to 0.05. Hence, the use of GAP layer, and possibly other

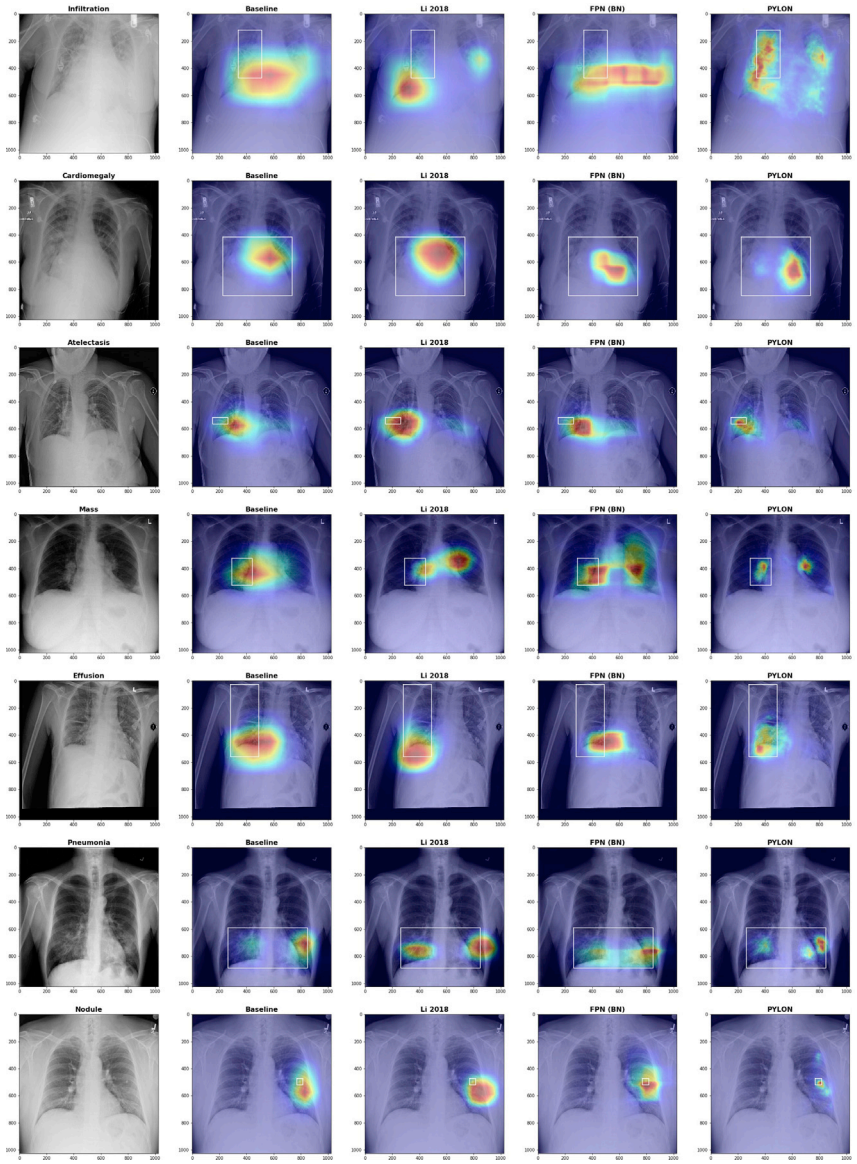


Figure 2. Example heatmaps generated from the NIH's Chest X-ray 14 dataset

Input image size was 256 × 256 pixels. White boxes denote ground truth annotations. The leftmost column shows the original images, and the other columns show the heatmaps from each model. Model names are indicated on top of each image. It should be noted that all models achieved similar IoU scores while producing qualitatively different heatmaps. More examples are provided in Figures S1 and S2. Example heatmaps for the larger input size of 512 × 512 pixels are provided in Figures S3 and S4.

global pooling layers that collapse spatial information, should be discouraged because it reduced the localization power of the resulting CAM heatmaps. Full results are provided in Table S18.

Another observation is that the original design of FPN (Kirillov et al., 2019) with group normalization yielded unintelligible heatmaps (Figure S14) while achieving a similar classification performance as other models. Replacing group normalization by batch normalization or setting the number of group parameter in group normalization to one, which makes it equivalent to batch normalization with the batch size of one, resolved the issue. Interestingly, the incompatibility between group normalization and CAM appeared to be specific to the architecture of FPN. Adding group normalization to PYLON or Li 2018 models only slightly reduced localization performances (Figure S15). It is unexpected for a normalization layer to have such a negative

Table 3. Point localization accuracy on the VinDr-CXR dataset

	Without transfer learning								With transfer learning		
Class	Avg. Area	Baseline	GradCAM[a]	GradCAM++[a]	XGradCAM[a]	Li et al. (2018b)	FPN	PYLON	Baseline	PYLON	PYLON two-phase
Weight. avg.	0.04	0.29	0.29	0.19	0.26	0.25	0.35	0.37	0.32	0.46	0.58
Macro avg.	0.06	0.31	0.31	0.2	0.27	0.29	0.36	0.42	0.35	0.49	0.56
Aortic enlargement	0.02	0.14	0.14	0.11	0.13	0.24	0.33	0.19	0.14	0.53	0.77
Atelectasis	0.07	0.35	0.35	0.26	0.29	0.35	0.39	0.45	0.51	0.48	0.71
Calcification	0.02	0.1	0.1	0.05	0.07	0.13	0.13	0.24	0.1	0.3	0.4
Cardiomegaly	0.07	0.76	0.76	0.46	0.65	0.34	0.7	0.68	0.83	0.61	0.79
Consolidation	0.05	0.56	0.56	0.28	0.45	0.56	0.6	0.77	0.52	0.79	0.82
ILD	0.16	0.54	0.54	0.38	0.41	0.62	0.62	0.72	0.63	0.75	0.78
Infiltration	0.06	0.44	0.44	0.28	0.35	0.41	0.56	0.71	0.51	0.75	0.74
Lung Opacity	0.05	0.36	0.36	0.23	0.33	0.43	0.42	0.55	0.37	0.62	0.64
Nodule/Mass	0.02	0.19	0.19	0.1	0.15	0.15	0.2	0.32	0.24	0.42	0.44
Other lesion	0.05	0.16	0.16	0.12	0.11	0.13	0.13	0.15	0.15	0.16	0.24
Pleural effusion	0.05	0.34	0.34	0.22	0.34	0.18	0.35	0.39	0.34	0.45	0.5
Pleural thickening	0.01	0.02	0.02	0.02	0.02	0.06	0.03	0.03	0.02	0.05	0.14
Pneumothorax	0.1	0.24	0.24	0.13	0.24	0.29	0.29	0.25	0.33	0.35	0.31
Pulmonary fibrosis	0.04	0.19	0.19	0.13	0.19	0.22	0.29	0.47	0.26	0.55	0.61

CIs are provided in supplemental information.
[a]Variation of GradCAM methods was performed on the Baseline model.

effect on CAM's heatmap quality while not impacting classification performance. From our analyses, we could not provide a satisfactory explanation for this phenomenon, and it would be an interesting avenue for further investigation.

DISCUSSION

In this study, we proposed PYLON that produces a more accurate localization explanation by increasing the resolution of the heatmaps produced by CAM methods. PYLON was compared to a wide range of CAM methods and previous works, including strong segmentation models like PAN (Li et al., 2018a), FPN (Kirillov et al., 2019), and DeeplabV3+ (Chen et al., 2018), in both general and medical image domains, namely the NIH's Chest X-ray 14, VinDR-CXR, and Pascal VOC2012 datasets. PYLON substantially improved CAM's heatmap accuracy across most classes in both general and medical image domains, especially for small object instances, without sacrificing much in terms of classification performance (Tables 1, 2, 3, 4, S7, S11, and S15). We also demonstrated a two-phase fine-tuning procedure that further improved the performance of PYLON on the small VinDR-CXR and Pascal VOC2012 datasets over the standard transfer learning technique (Tables 3 and 4). PYLON's ability to produce high-resolution, precise heatmaps that pinpoint the object locations is a strong explainability advantage because in critical applications such as abnormality classification from chest radiographs, the clinicians need to verify that the lesion is indeed present in the heatmap regions before accepting the classification outputs. Additionally, potential classification errors can be easily spotted when no object is found in the heatmap regions. In contrast, broad heatmaps derived from direct upscaling of low-resolution heatmaps do not allow the human operators to conclude whether the model made the correct classifications based on the pixels of the object or the model made erroneous classifications based on nearby unrelated pixels.

To measure the ability of CAM heatmap to direct human operators toward the object location, we proposed a new metric, named point localization accuracy, that focuses on the pixel with the highest score. In all experiments, point localization accuracy is more representative of the heatmap's explainability power than a standard object detection metric, such as IoU, which are roughly the same for all models. Nonetheless, PYLON still achieved higher IoU compared to the scores reported in previous works (Wang et al., 2017; Yao et al., 2018), even with the disadvantage of smaller input image size (Table

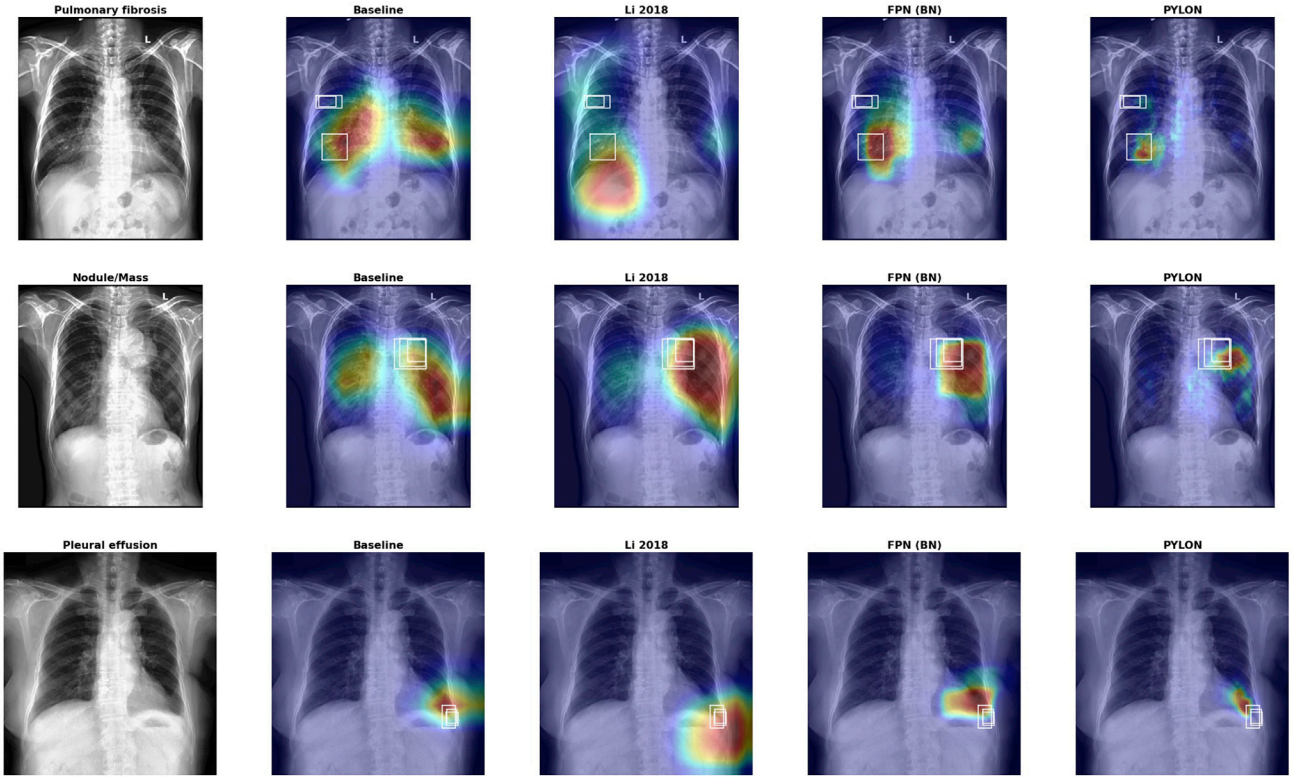


Figure 3. Example heatmaps generated from the VinDr-CXR dataset

White boxes denote ground truth annotations. The leftmost column shows the original images, and the other columns show the heatmaps from each model. Model names are indicated on top of each image. More examples are provided in Figures S5 and S6.

S19). Furthermore, point localization accuracy on classes with small object instances is also a good proxy for measuring the resolution of the heatmap because a low-resolution heatmap cannot distinguish pixels that were upscaled from the same low-resolution bin. In this regard, PYLON achieves the highest localization accuracy on all classes with small object instances (nodules in chest radiographs, bottle, TV monitor, and potted plant in general images). It should be noted that all the second-best models on these classes are based on image segmentation architectures which were inherently developed to produce high-resolution outputs. The performance gaps significantly widen if PYLON is compared to models based on classification architectures.

In addition to point localization accuracy, it is possible to assess the accuracy of heatmaps by comparing the model's predictions before and after masking out image regions with high heatmap scores. Because high-quality heatmaps should assign high scores to important regions of the images, removing such regions would have a strong impact on the model's behavior. Hence, models with higher heatmap quality would lose classification confidences more quickly than other models as more pixels with high heatmap scores were masked out. PYLON flipped its classification output from positive to negative when much smaller areas were masked out and achieved the best masked-out area on 12 out of 14 classes for the NIH's Chest X-ray 14 dataset and 14 out of 20 classes on the Pascal VOC2012 dataset (Tables S3, S4, and S13).

A key limitation of CAM is that when evidence for a certain class spans multiple sites across an image, such as opacity of the lung in a chest radiograph, CAM might focus on only one site (Figure 6). This is a well-known behavior of CAM methods (Huang et al., 2020; Bae et al., 2020) because the classification model can make an accurate prediction based on information from just the most discriminative part of the image. Hence, without extra knowledge on the number of sites or objects in an image, a model trained with only image-level annotation cannot reliably discover all occurrences of a given class. On the other hand, CAM also frequently highlights sensible regions that are not annotated by human experts. For example, in Figure 2 (Mass), almost all heatmaps highlighted both lungs while only one lung is

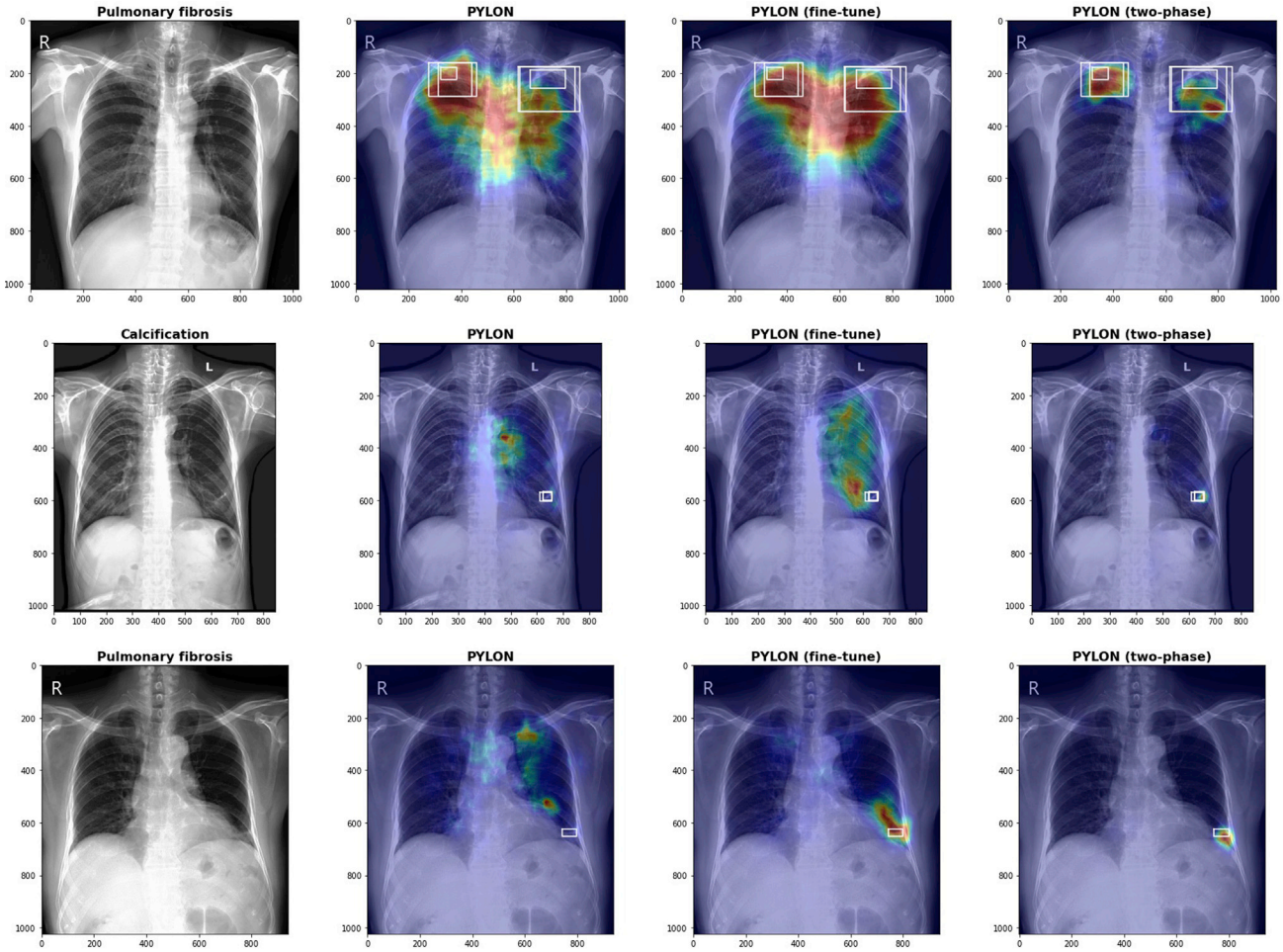


Figure 4. Impact of the two-phase fine-tuning technique on PYLON on VinDr-CXR dataset

White boxes denote ground truth annotations. The leftmost column shows the original images, and the other columns show the heatmaps from each model. Fine-tune indicates the standard transfer learning technique. Two-phase indicates our proposed two-phase fine-tuning technique. More examples are provided in Figures S7 and S8.

annotated. It is possible that CAM was confused by the presence of similar features in the other lung, but another interpretation is that of imperfect annotation. In the latter case, an accurate CAM can aid the discovery of unannotated objects.

CAM's heatmaps can also point to nonsensical locations. Clear examples are FPN's heatmaps before the group norm issue was resolved (Figure S14). More subtle examples can be found in Figure S4 (Mass) where the highlighted areas are dispersed across the lungs while a mass object is expected to be highly localized. There are multiple possible explanations. Low confidence in classification prediction is the most likely cause because CAM relies heavily on the classifier. If the classifier is uncertain about a particular class, the heatmap for that class is likely to be of low quality. Another explanation is on the deep learning model architecture. Not all designs are suitable for CAM methods. Substituting group norm with batch norm in FPN led to a drastic improvement in CAM's accuracy (Figure S14). The use of global average pooling (GAP) was also observed to degrade the quality of CAM heatmaps.

A much larger dataset is required to achieve high-quality heatmaps than what is needed to gain high classification performance. While the size of 15,000 images of VinDr-CXR dataset is enough to achieve stellar classification performances (Table S11, >0.95 macro average classification AUROC), the resulting heatmaps were much worse compared to those of the NIH's dataset (Figures 2 and 3, macro average point localization accuracy of 0.46 versus 0.60) and generally missed the annotated bounding boxes. This issue

Table 4. Point localization accuracy on the Pascal VOC2012 dataset

Class	Avg.Area	Baseline	GradCAM[a]	GradCAM++[a]	XGradCAM[a]	Li et al. (2018b)	FPN	PYLON	PYLON two-phase
Weighted. avg.	0.33	0.77	0.77	0.59	0.74	0.45	0.81	0.78	0.83
Macro avg.	0.35	0.77	0.77	0.58	0.72	0.51	0.8	0.77	0.83
Aeroplane	0.4	0.85	0.85	0.69	0.83	0.5	0.88	0.84	0.9
Bicycle	0.37	0.85	0.85	0.62	0.8	0.62	0.85	0.87	0.89
Bird	0.26	0.77	0.77	0.57	0.74	0.27	0.83	0.75	0.86
Boat	0.28	0.52	0.52	0.41	0.48	0.36	0.57	0.51	0.5
Bottle	0.13	0.4	0.4	0.27	0.36	0.21	0.47	0.45	0.56
Bus	0.44	0.87	0.87	0.72	0.82	0.68	0.91	0.9	0.92
Car	0.26	0.63	0.63	0.48	0.6	0.31	0.7	0.67	0.75
Cat	0.52	0.95	0.95	0.79	0.92	0.57	0.95	0.92	0.97
Chair	0.26	0.62	0.62	0.41	0.54	0.47	0.62	0.58	0.66
Cow	0.39	0.88	0.88	0.61	0.78	0.68	0.88	0.85	0.92
diningtable	0.33	0.82	0.82	0.53	0.74	0.72	0.84	0.83	0.81
Dog	0.42	0.91	0.91	0.69	0.87	0.56	0.92	0.92	0.94
horse	0.42	0.91	0.91	0.72	0.85	0.62	0.92	0.89	0.93
motorbike	0.43	0.86	0.86	0.7	0.83	0.63	0.87	0.85	0.91
person	0.36	0.81	0.81	0.65	0.81	0.32	0.86	0.86	0.88
pottedplant	0.22	0.58	0.58	0.36	0.52	0.47	0.66	0.63	0.73
sheep	0.34	0.79	0.79	0.6	0.74	0.46	0.85	0.8	0.89
sofa	0.48	0.85	0.85	0.62	0.77	0.71	0.87	0.82	0.86
train	0.42	0.86	0.86	0.7	0.83	0.65	0.77	0.82	0.86
tvmonitor	0.2	0.66	0.66	0.44	0.61	0.34	0.72	0.62	0.79

CIs are reported in supplemental information.

[a]Variation of GradCAM methods was performed on the Baseline model. IoU cannot be fairly calculated in these cases.

can be largely alleviated via our proposed two-phase fine-tuning technique (Figure 4), which improved the macro average point localization accuracy for the VinDr-CXR dataset from 0.46 to 0.58 and from 0.78 to 0.83 for Pascal VOC2012 (Tables 3 and 4).

Although global average pooling (GAP) can 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 GAP decreases the accuracy of CAM heatmaps. Removal of GAP from PAN (Li et al., 2018a) and DeeplabV3+ (Chen et al., 2018) resulted in more stable and higher-quality heatmaps (Table S18). This is expected because GAP, by design, collapses information across spatial dimensions which are necessary for reconstructing accurate heatmaps in downstream layers. However, because most deep models contain multiple paths for information to propagate through, some spatial information may still be preserved through paths that do not contain global pooling. This might explain the high variance in heatmap quality across experiments for models with GAP as the propagation of spatial information through these networks would be highly stochastic. In conclusion, we recommend against using GAP in PAN-like decoder architecture.

Interpretation of CAM should always be mindful of the fact that CAM methods explain only a small downstream portion of a deep model. By only explaining the high-level decision making, CAM heatmaps are easily interpreted by human operators. While this is the main strength of CAM methods, one should be aware that regions highlighted on the heatmaps are not the only parts of the image that the model considered. This is evident in Cardiomegaly class (Figure 2) where the heatmaps conveniently highlight just the heart while leaving out the perimeter of the thorax that is also necessary for defining cardiomegaly. Hence, while CAM may be effective for pinpointing the location of objects in an image, it may be incapable of explaining more sophisticated decisions.

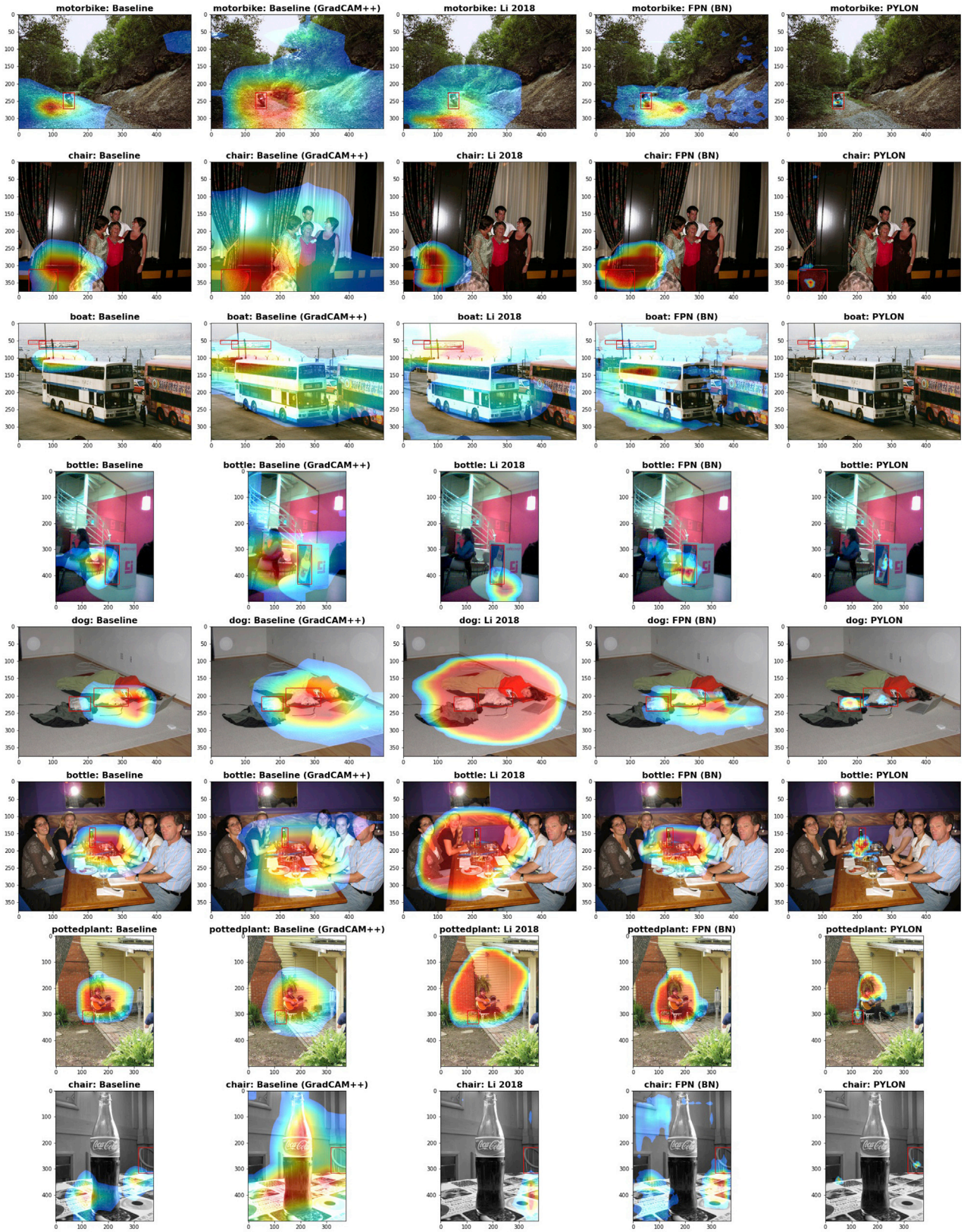

motorbike: Baseline
motorbike: Baseline (GradCAM++)
motorbike: Li 2018
motorbike: FPN (BN)
motorbike: PYLON
chair: Baseline
chair: Baseline (GradCAM++)
chair: Li 2018
chair: FPN (BN)
chair: PYLON
boat: Baseline
boat: Baseline (GradCAM++)
boat: Li 2018
boat: FPN (BN)
boat: PYLON
bottle: Baseline
bottle: Baseline (GradCAM++)
bottle: Li 2018
bottle: FPN (BN)
bottle: PYLON
dog: Baseline
dog: Baseline (GradCAM++)
dog: Li 2018
dog: FPN (BN)
dog: PYLON
bottle: Baseline
bottle: Baseline (GradCAM++)
bottle: Li 2018
bottle: FPN (BN)
bottle: PYLON
pottedplant: Baseline
pottedplant: Baseline (GradCAM++)
pottedplant: Li 2018
pottedplant: FPN (BN)
pottedplant: PYLON
chair: Baseline
chair: Baseline (GradCAM++)
chair: Li 2018
chair: FPN (BN)
chair: PYLON

Figure 5. Example heatmaps generated from the Pascal VOC2012 dataset

Red boxes denote ground truth annotations. Each column shows the heatmaps from each model. Model names are indicated on top of each image. More examples are provided in Figures S9–S12.

Finally, as a cautionary note, CAMs, like other aspects of deep learning, are susceptible to biases in real-world datasets. Deep models can easily exploit spurious correlations, such as the presence of medical tubes and lines in chest radiographs of patients with pneumothorax. Furthermore, medical data collected from different populations or geographical regions would have different disease prevalence that could be memorized as prior information by the models. Most importantly, data collection is always subjected to differences in labeling standards across experts and methods. For example, the NIH's Chest X-ray 14 dataset was mostly labeled by automatically extracting keywords from radiologist's reports, while the VinDR-CXR dataset was manually annotated by a panel of radiologists. Large-scale datasets such as NIH's Chest X-ray 14 dataset also consists of chest radiograph spanning multiple years during which the standards of diagnosis and data acquisition techniques could have changed.

Limitation of the study

A major limitation of this study is the lack of user validation that would confirm the practical impact of getting higher accuracy heatmaps from PYLON. We originally planned to enlist radiologists to score the quality of heatmaps on chest radiograph images but due to a combination of COVID-19 situation and the fact that radiologists also consider other clinical information when making diagnoses, the idea was placed on hold. Nonetheless, because images of general objects from the Pascal VOC2012 dataset are easy to understand by human eye, the increased quality of heatmaps produced by PYLON can still be appreciated even without expert validation. Another limitation is the lack of definitive metric for the explanatory power of heatmap. We had to invent point localization accuracy, evaluate intersection-over-union and masked-out area metrics, and show many examples of heatmaps to argue that PYLON improves the explainability of deep classification models.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCE TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Datasets
  - Performance metrics
  - Proposed model
  - Effective two-phase transfer learning approach
  - Benchmark models
  - Ablation analyses
  - Hyperparameter tuning
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.103933.

ACKNOWLEDGMENTS

This work was supported by the Grant for Special Task Force for Activating Research, Ratchadapisek Sompoch Endowment Fund, Chulalongkorn University (to E.C. and S.S.). We also thank CMKL University, Thailand, for computing quotas on their DGX-Pod.

AUTHOR CONTRIBUTIONS

Conceptualization, all authors; methodology and experimental designs, all authors; model development, K.P.; data analysis, K.P.; writing – original draft preparation, K.P.; writing – review and editing, all authors; computational resources, B.K.

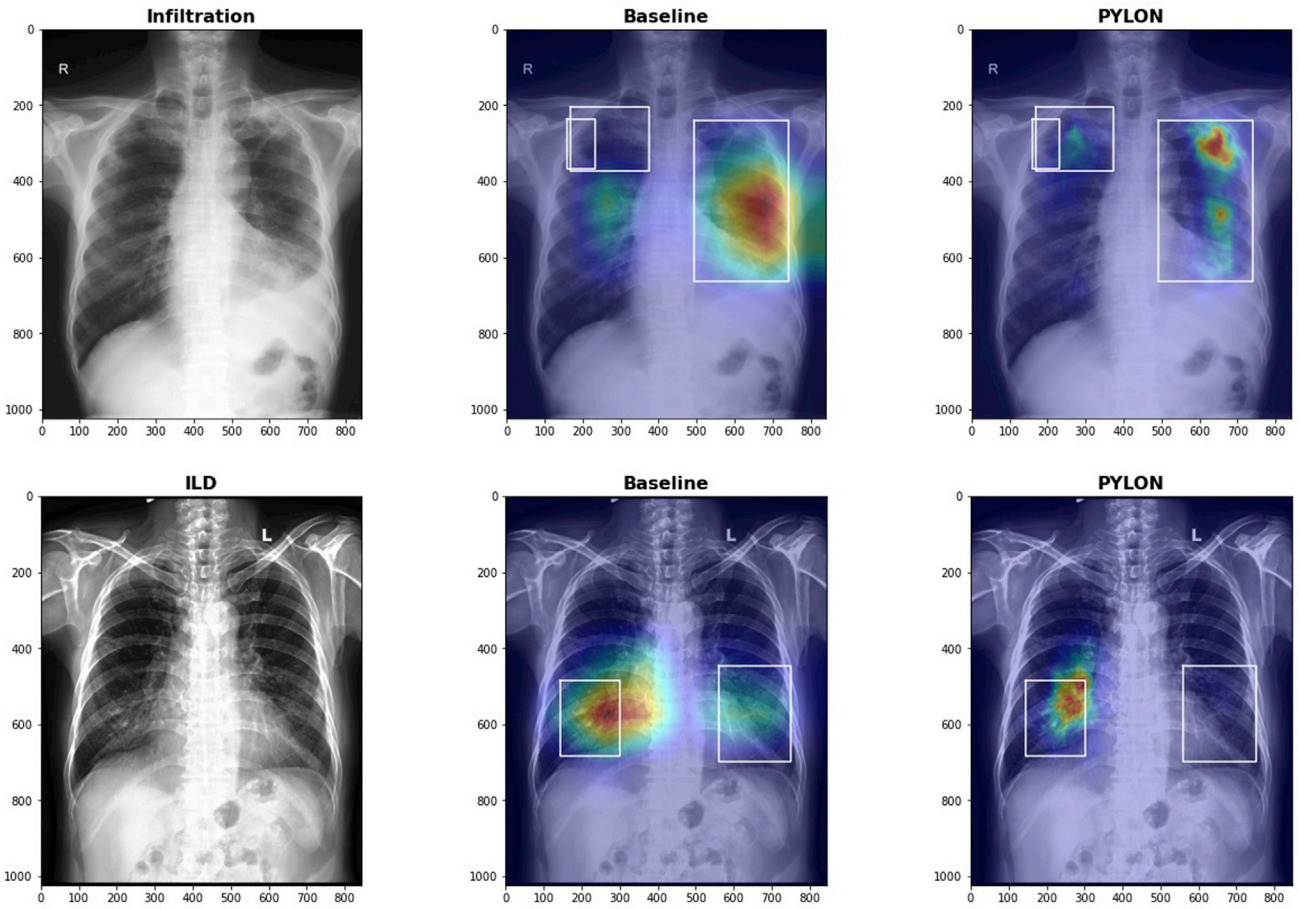


Figure 6. CAM cannot reliably discover all objects

As detecting one object is sufficient for the model to make the correct classification, CAM heatmap may highlight only one object instance in an image with multiple objects. Example images were taken from the VinDR-CXR dataset.

DECLARATION OF INTERESTS

The authors declare no competing interest.

Received: November 13, 2021
Revised: January 27, 2022
Accepted: February 11, 2022
Published: March 18, 2022

REFERENCES

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) (IEEE), pp. 2204–2213.

Ahn, J., and Kwak, S. (2018). Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. Preprint at arXiv. http://arxiv.org/abs/1803.10464.

Ancona, M., Ceolini, E., Öztireli, C., and Gross, M. (2017). Towards better understanding of gradient-based attribution methods for deep neural networks. Preprint at arXiv. http://arxiv.org/abs/1711.06104.

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*, e0130140.

Bae, W., Noh, J., and Kim, G. (2020). Rethinking class Activation mapping for weakly supervised object localization. In Computer Vision – ECCV 2020 (Springer International Publishing), pp. 618–634, Lecture Notes in Computer Science.

Chang, C.H., Creager, E., Goldenberg, A., and Duvenaud, D. (2018). Explaining image classifiers by counterfactual generation. Preprint at arXiv. http://arxiv.org/abs/1807.08024.

Chattopadhyay, A., Sarkar, A., Howlader, P., and Balasubramanian, V.N. (2017). Grad-CAM++: improved visual explanations for deep convolutional networks. Preprint at arXiv. http://arxiv.org/abs/1710.11063.

Chen, L.C., George, P., Schroff, F., and Adam, H. (2017a). Rethinking atrous convolution for semantic image segmentation. Preprint at arXiv. http://arxiv.org/abs/1706.05587.

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) (IEEE), pp. 5659–5667.

Chen, L.C., Zhu, Y., George, P., Schroff, F., and Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In Computer Vision – ECCV 2018 (Springer International Publishing), pp. 833–851.

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) (IEEE), pp. 2214–2223.

Cinbis, R.G., Verbeek, J., and Schmid, C. (2017). Weakly supervised object localization with multi-fold multiple instance learning. IEEE Trans. Pattern Anal. Mach. Intell. *39*, 189–203.

Desai, S., and Ramaswamy, H.G. (2020). Ablation-CAM: visual explanations for deep convolutional network via gradient-free localization. In 2020 IEEE Winter Conference on Applications of Computer Vision (WACV) (IEEE), pp. 972–980.

Dietterich, T.G., Lathrop, R.H., and Lozano-Pérez, T. (1997). Solving the multiple instance problem with axis-parallel rectangles." artificial intelligence. https://doi.org/10.1016/s0004-3702(96)00034-3.

Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., and Zisserman, A. (2011). The pascal visual object classes challenge 2012 (Voc2012) results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

Fan, L., Zhao, S., and Ermon, S. (2017). Adversarial localization network. In Learning with Limited Labeled Data: Weak Supervision and beyond, NIPS Workshop. https://www.semanticscholar.org/paper/Adversarial-Localization-Network-Fan/4e0636a1b92503469b44e2807f0bb35cc0d97652.

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

Fu, R., Hu, Q., Dong, X., Guo, Y., Gao, Y., and Li, B. (2020). Axiom-based grad-CAM: towards accurate visualization and explanation of CNNs. Preprint at arXiv. http://arxiv.org/abs/2008.02312.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. Preprint at arXiv. http://arxiv.org/abs/1502.01852.

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

He, K., Zhang, X., Ren, S., and Sun, J. (2016a). Identity mappings in deep residual networks. In *Computer Vision – ECCV 2016*, 9908 LNCS:630–45 (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 (IEEE), 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 (IEEE), pp. 7014–7023.

Huang, Z., Zou, Y., Bhagavatula, V., and Huang, D. (2020). Comprehensive attention self-distillation for weakly-supervised object detection. Preprint at ArXiv [Cs.CV]. http://arxiv.org/abs/2010.12023.

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*), pp. 448–456, ICML'15. JMLR.org.

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., et al. (2019). CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. Proc. AAAI Conf. Artif. I. *33*, 590–597.

Kervadec, H., Jose, D., Tang, M., Granger, E., Boykov, Y., and Ismail Ben Ayed. (2018). Constrained-CNN losses for weakly supervised segmentation. Med. Image Anal. https://doi.org/10.1016/j.media.2019.02.009.

Kervadec, H., Jose, D., Wang, S., Granger, E., and Ayed, I.B. (2020). Bounding boxes for weakly supervised segmentation: global constraints get close to full supervision. Preprint at arXiv. http://arxiv.org/abs/2004.06816.

Kingma, D.P., and May, J.B. (2015). Adam: a method for stochastic optimization. In 3rd International Conference on Learning Representations, ICLR 2015. https://dblp.org/rec/journals/corr/KingmaB14.html.

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)* (IEEE), pp. 6392–6401.

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

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*). https://www.semanticscholar.org/paper/Pyramid-Attention-Network-for-Semantic-Segmentation-Li-Xiong/bea705a6a3793ffa2f61c33a82b759ad9f706947.

Li, Z., Wang, C., Han, M., Xue, Y., Wei, W., Li, L., and Fei-Fei, L. (2018b). 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)* (IEEE), pp. 936–944.

Lin, Z.Q., Javad Shafiee, M., Bochkarev, S., Jules, M.S., Wang, X.Y., and Wong, A. (2019). Do explanations reflect decisions? A machine-centric strategy to quantify the performance of explainability algorithms. Preprint at arXiv. http://arxiv.org/abs/1910.07387.

Liu, J., Zhao, G., Fei, Y., Zhang, M., Wang, Y., and Yu, Y. (2019). 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) (IEEE), pp. 10631–10640.

Liu, Y., Wu, Y.H., Wen, P.S., Shi, Y.J., Qiu, Y., and Cheng, M.M. (2020). Leveraging instance-, image- and dataset-level information for weakly supervised instance segmentation. IEEE Trans. Pattern Anal. Mach. Intell. https://doi.org/10.1109/TPAMI.2020.3023152.

Mopuri, K.R., Garg, U., and Venkatesh Babu, R. (2019). CNN fixations: an unraveling approach to visualize the discriminative image regions. IEEE Trans. Image Process. A Publ. IEEE Signal Process. Soc. *28*, 2116–2125.

Muhammad, M.B., and Yeasin, M. (2020). Eigen-CAM: class Activation map using principal components. Preprint at arXiv. http://arxiv.org/abs/2008.00299.

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., et al. (2020). VinDr-CXR: an open dataset of chest X-rays with radiologist's annotations. Preprint at arXiv. http://arxiv.org/abs/2012.15029.

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) (IEEE), pp. 685–694.

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., et al. (2017). CheXNet: radiologist-level pneumonia detection on chest X-rays with deep learning. Preprint at arXiv, 1711.05225.

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). https://doi.org/10.1109/cvpr42600.2020.01061.

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 (Springer International Publishing), pp. 234–241.

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

Russakovsky, O., Jia, D., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., et al. (2015). ImageNet large scale visual recognition challenge. Int. J. Computer Vis. *115*, 211–252.

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) (IEEE), pp. 618–626.

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, *70*Proceedings of the 34th International Conference on Machine Learning (The Proceedings of Machine Learning Research Press), pp. 3145–3153, ICML'17. JMLR.org.

Sim, Y., Chung, M.J., Kotter, E., Yune, S., Kim, M., Do, S., Han, K., et al. (2020). Deep convolutional neural network-based software improves radiologist detection of malignant lung nodules on chest radiographs. Radiology *294*, 199–209.

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

Simonyan, Karen, and Zisserman, Andrew (2015). Very deep convolutional networks for large-scale image recognition. In 3rd International Conference on Learning Representations, *ICLR 2015*, Yoshua Bengio and Yann LeCun, eds.. https://www.semanticscholar.org/paper/eb42cf88027de515750f230b23b1a057dc782108.

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 Proceedings of the 36th International Conference on Machine Learning, PMLR, *97*, K. Chaudhuri and R. Salakhutdinov, eds (The Proceedings of Machine Learning Research Press), pp. 6105–6114.

Wang, H., Wang, Z., Du, M., Fan, Y., Zhang, Z., Ding, S., Mardziel, P., and Hu, X. (2019). Score-CAM: score-weighted visual explanations for convolutional neural networks. Preprint at arXiv. http://arxiv.org/abs/1910.01279.

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) (IEEE), pp. 3462–3471.

Wei, Y., Feng, J., Liang, X., Cheng, M.M., Zhao, Y., and Yan, S. (2017). Object region mining with adversarial erasing: a simple classification to semantic segmentation approach. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (IEEE), pp. 1568–1576.

Yang, K., Li, D., and Dou, Y. (2019). Towards precise end-to-end weakly supervised object detection network. Preprint at arXiv. http://arxiv.org/abs/1911.12148.

Yao, L., Jordan, P., Poblenz, E., Ben, C., and Lyman, K. (2018). Weakly supervised medical diagnosis and localization from multiple resolutions. Preprint at arXiv. http://arxiv.org/abs/1803.07703.

Zeiler, M.D., and Rob, F. (2014). Visualizing and understanding convolutional networks. In Computer Vision – ECCV 2014 (Springer International Publishing), pp. 818–833.

Zhang, J., Bargal, S.A., Lin, Z., Brandt, J., Shen, X., and Sclaroff, S. (2018a). Top-down neural attention by excitation backprop. Int. J. Computer Vis. *126*, 1084–1102.

Zhang, R. (2019). Making convolutional networks shift-invariant again. In Proceedings of the 36th International Conference on Machine Learning, PMLR. http://proceedings.mlr.press/v97/zhang19a.html.

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 (IEEE), 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) (IEEE), 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) (IEEE), pp. 1859–1868.

STAR★METHODS

KEY RESOURCE TABLE

REAGENT OR RESOURCE	SOURCE	IDENTIFIER
Python	https://www.python.org/	Version 3.7
Albumentations	https://albumentations.ai/	Version 0.5.2
Pandas	https://pandas.pydata.org/	Version 1.1.3
Matplotlib	https://matplotlib.org/	Version 3.3.1
PyTorch	https://pytorch.org/	Version 1.7.1
Torchvision	https://pytorch.org/vision/	Version 0.11.0
CUDA Toolkit	https://developer.nvidia.com/cuda-toolkit	Version 11.0

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and/or reagents should be directed to and will be fulfilled by Ekapol Chuangsuwanich (ekapol.c@chula.ac.th)

Materials availability

This study did not generate new reagents.

Data and code availability

Code: The source codes for reproducing all the main experiments (not including the ablation studies) are available on GitHub at https://github.com/cmbchula/pylon.

Dataset: All datasets used in this work are publicly available. NIH's Chest X-ray 14 dataset is available at https://nihcc.app.box.com/v/ChestXrayNIHCC. VinDr-CXR dataset (Kaggle version) is available at https://www.kaggle.com/c/vinbigdata-chest-xray-abnormalitiesdetection/overview. Pascal VOC 2012 dataset is available at http://host.robots.ox.ac.uk/pascal/VOC/voc2012/.

Additional information: Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Datasets

We conducted experiments on medical image and general image domains. For medical image, two chest radiograph datasets, NIH's Chest X-ray 14 (Wang et al., 2017) and VinDr-CXR (Nguyen et al., 2020) were used. For general image, Pascal VOC2012 (Everingham et al., 2011) was used. All datasets contain region-level annotation which is required for assessing the accuracy of class-activation map (CAM)'s heatmaps. The NIH's Chest X-ray 14 dataset contains more than 100,000 frontal (AP and PA) chest radiographs. Image-level classification labels for 14 chest abnormalities were automatically extracted from radiologist reports. Only 880 images (across 8 abnormalities) were manually annotated with bounding boxes by board-certified radiologists. The VinDr-CXR dataset is much smaller with 18,000 PA chest radiographs (15,000 available for training at the time of writing), all of which were annotated with bounding boxes across 15 abnormalities by three different radiologists each with at least 8 years of experience (from a pool of 17 radiologists). It should be noted that the Kaggle competition version of this dataset was used here. The Pascal VOC2012 dataset contains 11,530 realistic scenes annotated with 20 object classes. Each scene is annotated with bounding boxes, and some are also annotated with segmentation masks. This dataset is widely used to benchmark object detection and weakly-supervised learning models (Ahn and Kwak, 2018; Liu et al., 2020; Huang et al., 2020). In all experiments, bounding box annotations were used only for evaluation and not for training the models.

Performance metrics

Three metrics, point localization accuracy, masked-out area, and intersection over union (IoU), were used to assess the quality and accuracy of the CAM heatmaps produced by each model. The description, advantages, and disadvantages of each metric are outlined below:

Point localization accuracy of a class *C* is defined as the fraction of images labeled with class *C* where the highest-valued pixels of the corresponding heatmaps are located within the ground truth bounding boxes of class *C*. This was selected as the main metric in our analyses because point localization accuracy does not assume that each bounding box contains the whole object and only the object itself. Furthermore, point localization accuracy reflects real-world practices of pointing an arrow to the object location, as done by radiologists. This metric may be under-specific because other pixels of the heatmap were not considered in the calculation.

Masked-out area is the average size of image area that can be masked-out (e.g., replaced by zeros, sorted according to the heatmap intensity) before a deep model switches its prediction for the correct class from positive to negative. This is implemented by gradually increasing the masked-out area by 0.5% of the input image area per step until the prediction flipped. There are many related works on this technique including adversarial erasing (Lin et al., 2019; Fan et al., 2017; Chang et al., 2018; Zhang et al., 2018b; Ren et al., 2020; Wei et al., 2017; Choe and Shim 2019; Fong and Vedaldi, 2017). The downside of this metric is that it is computational expensive and has high variance, making the comparison across models unreliable. Moreover, once a large portion of the input image is masked, the input image can be so different from the training distribution that the deep model's behavior becomes unpredictable. Masked-out areas are reported in full in Supplemental Information.

Intersection over Union (IoU) is often used to measure the closeness between the generated heatmap and human-annotated ground truth. This metric is appropriate for object detection where the exact boundary of an object matters. In the context of explaining deep model's decision, IoU is overly specific because it favors heatmaps that cover the whole bounding box, including pixel regions that may be irrelevant for the classification. Our results showed that comparable IoU scores were achieved by models with widely different heatmap qualities, while point localization accuracy better represents the heatmap quality.

It should be noted that a classification threshold (for predicting positive and negative) needs to be set for masked-out area and IoU. To ensure a fair selection, the threshold for each method is selected automatically to achieve the maximal geometric mean of specificity and sensitivity. Therefore, it is impossible to set a threshold for CAM methods whose heatmap intensities do not represent the actual prediction confidence. These include GradCAM (Selvaraju et al., 2017), GradCAM++ (Chattopadhyay et al., 2017), and XGradCAM (Fu et al., 2020).

While improving the classification performance is not a main focus of this work, the areas under the receiver operating characteristic curve (AUROC) are also calculated to ensure that improvement in heatmap accuracy does not come at a significant reduction in classification performance.

Proposed model

Pyramid Localization Network (PYLON). Low-resolution heatmap is arguably the main hindrance of CAM methods that limits their explainability power. But this issue has been overlooked in recent works (Chattopadhyay et al., 2017; Wang et al., 2019; Muhammad and Yeasin, 2020; Desai and Ramaswamy 2020; Fu et al., 2020). In this work, we designed PYLON as a general principle for extending a deep classifier to increase the resolution of heatmaps produced by CAM methods with minimal computational cost. PYLON consists of three parts: an encoder, a decoder, and a prediction head (Figure 1B). The encoder can be any standard deep classification model, such as ResNet (He et al., 2016b), DenseNet (Huang et al., 2017), VGG (Simonyan and Zisserman 2015), and EfficientNet (Tan and Le 2019). The decoder is composed of two modules, Pyramid Attention module (PA) and Upsampling module (UP).

Pyramid attention module (PA) is a form of spatial attention that was found to be 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 that utilizes different input scales to handle inputs of varying sizes (Lin et al., 2017; Chen et al., 2017a; Kirillov et al., 2019; Li et al., 2018a; Chen et al., 2018). The PA

module compresses its input signals into a single channel while consecutively down-scaling it to varying degrees. At each scale, the signal is passed through a non-linear transformation and up-scaled back for the final combination across scales into a spatial attention mask which has one channel. This spatial attention mask was then multiplied to the transformed encoder feature maps like any other attention mask. The PA module takes the uppermost 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 UPs modules.

Upsampling module (UP) is a lightweight two-layer 1 × 1 convolution with batch norm (Ioffe and Szegedy 2015) and ReLU activation from the lateral skip connection from the corresponding block in the encoder. The output of this is added to the bilinearly upscaled (2× horizontally and 2× vertically) decoded feature maps from the PA module or the output of the previous UP module. There can be as many UP modules as the number of blocks in the encoder minus one. A block of layers in this context is a group of layers that operates on the same resolution. For example, there are 4 blocks in ResNet and its variants.

Prediction head combines the output of the last UP module by a linear function (equivalent to Conv 1 × 1) for each class to get the corresponding heatmaps. To train with classification loss, each class's heatmap is passed through a global max pooling layer to get the prediction. The training of PYLON is the same as a regular classification network except that high accuracy heatmaps are produced as by-product.

PYLON is simple and fast by nature as it relies mostly on Conv 1 × 1 which has small computational cost and memory overhead. PYLON explains the encoder's prediction by refining the encoder's low-resolution heatmaps with signals from the encoder's shallower layers to generate high-resolution heatmaps. Since PYLON relies on the encoder entirely for classification, any improvement on the encoder also improves PYLON as well (Simonyan and Zisserman 2015; Krizhevskyet al., 2012; Huang et al., 2017; Hu et al., 2018; Tan and Le 2019; He et al., 2016a, He et al., 2016b).

Effective two-phase transfer learning approach

Transfer learning has become a standard practice to train neural nets on more limited datasets. When using PYLON, 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. This phase effectively trains PYLON's decoder and prediction head without disrupting the encoder. 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 much better retains localization knowledge from the source dataset than the standard transfer learning procedure does.

Benchmark models

As the primary objective of our work is to explain deep model's decision via heatmaps, related works in weakly-supervised object detection (Zhu et al., 2017; Yang et al., 2019; Huang et al., 2020; Ren et al., 2020), which aim to produce bounding boxes that cover the entire object, and works in weakly-supervised semantic segmentation (Kervadec et al., 2018; Huang et al., 2018; Ahn et al., 2019; Kervadec et al., 2020), which aim to classify every pixel in the input, were not considered. Additionally, works that employed non-heatmap explanation methods (Simonyan et al., 2013; Zeiler and Rob, 2014; Bach et al., 2015; Shrikumar, Greenside, and Kundaje 2017) were also excluded.

PYLON was compared against other methods in the CAM family, including the original CAM (Oquab et al., 2015; Zhou et al., 2016) (also referred to as Baseline), GradCAM (Selvaraju et al., 2017), GradCAM++ (Chattopadhyay et al., 2017), and XGradCAM (Fu et al., 2020). Both GradCAM and GradCAM++ did not perform well on binary classification task with sigmoid activation due to their application of ReLU on the heatmaps and gradients. Hence, we used GradCAM and GradCAM++ without ReLU in this work to achieve better results. ScoreCAM (Wang et al., 2019) and AblationCAM (Desai and Ramaswamy 2020) were excluded due to their very high computational cost. EigenCAM (Muhammad and Yeasin, 2020) was excluded because it does not produce a class-specific heatmap. Each CAM method was applied to the same ResNet-50 classification model with the same input image size and feature map resolution to ensure a fair comparison.

Since the main contribution of PYLON is the high-resolution heatmaps, we also compared PYLON with other high-resolution models (with the same ResNet-50 as encoder), including Li 2018 (Li et al., 2018b) which proposed both a model and a multi-instance loss function (MIL), and major image segmentation models such as Unet (Ronneberger et al., 2015), FPN (Kirillov et al., 2019) with batch normalization (BN) or group normalization, PAN (Li et al., 2018a), and DeeplabV3+ (Chen et al., 2018). These models have an innate ability to produce high accuracy heatmaps due to their high-resolution outputs. Each image segmentation model was transformed into a classification model by adding a global max pooling layer on top of its final output. Our implementation of Unet used a fairly larger number of channels of (256, 128, 64, 64, 64) than the default of (256, 128, 64, 32, 16).

There are related works that could not be compared directly because there was not enough information for us to faithfully re-implement their approaches. Wang 2017 (Wang et al., 2017) proposed a typical CAM model coupled with log-sum-exp pooling (LSE pool) which should be well represented by our Baseline model. Yao 2018 (Yao et al., 2018) proposed a model with high-resolution CAM and parameterized LSE-LBA pool. Instead, we applied the same performance metrics used in these works on PYLON and compared the results with previously reported scores. Rozenberg 2020 (Rozenberg et al., 2020) proposed a more numerically stable multi-instance loss function with some architectural improvements (Zhang 2019; Su et al., 2019), but it is not comparable to our study since it requires bounding box supervision during training.

Ablation analyses

Ablation analyses were performed on a small variant of PYLON (Figure S13) which has a single Conv 1 × 1 in its UP module instead of the two Conv 1 × 1 as in the proposed PYLON (Figure 1B). All three components of PYLON, PA module, UP module, and prediction head, were evaluated. For the PA module, the pyramidal attention path was removed while keeping the main Conv 1 × 1 in the PA module intact. For the UP module, both the design of UP module and the number of UP modules were investigated. We compare three different options of convolution in the UP module: a single-layer Conv 1 × 1 (small), a two-layer Conv 1 × 1 (PYLON), and a single-layer Conv 3 × 3 which is a common kernel size for convolution. We also compared PYLON with 0, 1, 2, or 3 UP modules. For the prediction head, we compare the proposed single-layer Conv 1 × 1 against a larger Conv 3 × 3.

To test whether PYLON's use of a heatmap output with a simple global maxpooling for classification output has an adverse effect on classification performance, we replaced the prediction head by a standard global average pooling followed by a multilayer perceptron with ReLU (which is the common architecture for classification models). The NIH's Chest X-ray 14 dataset was used for evaluation.

Hyperparameter tuning

All trainings used Adam optimizer (Kingma and May 2015) with a learning rate of 10^{-4} and without weight decay. The learning rate is reduced by 5× when the loss on the validation dataset does not improve for two consecutive epochs. The training was stopped when the learning rate was reduced by more than two times. The best model checkpoint was selected based on the loss on the validation set and subsequently evaluated on the test set. All images were resized to 256 × 256 (or 512 × 512) before feeding into the models.

QUANTIFICATION AND STATISTICAL ANALYSIS

Each experiment was repeated five times using different weight initializations to estimate the variances of model performances. Confidence intervals were calculated with Student's t distribution assumption (with n = 5) and reported in the Supplemental Information.