A Localisation Study of Deep Learning Models for Chest X-ray Image Classification

James Gascoigne-Burns
Department of Computer Science
Durham University
Durham, UK
james.gascoigne-burns@durham.ac.uk

Stamos Katsigiannis
Department of Computer Science
Durham University
Durham, UK
stamos.katsigiannis@durham.ac.uk

***Abstract*—Deep learning models have demonstrated superhuman performance in a multitude of image classification tasks, including the classification of chest X-ray images. Despite this, medical professionals are reluctant to embrace these models in clinical settings due to a lack of interpretability, citing being able to visualise the image areas contributing most to a model's predictions as one of the best ways to establish trust. To aid the discussion of their suitability for real-world use, in this work, we attempt to address this issue by conducting a localisation study of two state-of-the-art deep learning models for chest X-ray image classification, ResNet-38-large-meta and CheXNet, on a set of 984 radiologist annotated X-ray images from the publicly available ChestX-ray14 dataset. We do this by applying and comparing several state-of-the-art visualisation methods, combined with a novel dynamic thresholding approach for generating bounding boxes, which we show to outperform the static thresholding method used by similar localisation studies in the literature. Results also seem to indicate that localisation quality is more sensitive to the choice of thresholding scheme than the visualisation method used, and that a high discriminative ability as measured by classification performance is not necessarily sufficient for models to produce useful and accurate localisations.**

***Index Terms*—Deep Learning, Deep Learning Interpretability, Chest X-ray, Computer-Aided Diagnosis**

I. Introduction

Thoracic diseases are one of the leading causes of death globally [1], making chest X-ray (CXR) imaging the most common medical imaging exam in the world [2]. The ability to accurately identify abnormalities from CXRs is critical in being able to detect, treat and manage diseases, but even when adopting systematic approaches to diagnosis, radiologists can miss up to 15% of cases [3], either because of the very small size of pathological features [4], or because of fatigue that has been shown to occur after as little as an hour of continuous reading [3]. Following the release of the seminal ChestX-ray14 dataset [5], deep learning approaches have been proposed with diagnostic capabilities similar to those of practising radiologists [6], sparking interest in the use of deep learning models in computer-aided diagnosis (CAD) systems. Such CAD systems could help to guide radiologists' attention when taking readings or provide a second opinion, increasing the accuracy of diagnoses and helping to detect abnormalities sooner, leading to better preventive care and patient outcomes.

Interviewed radiologists have described the potential utility of localisations for detecting early-stage pathologies [7], where localisation refers to a model's ability to correctly identify the image area indicating a given class. However, despite its recognised importance, there is often no quantitative procedure carried out to measure a model's localisation performance, even among the state of the art [6], [8]. With the original release of ChestX-ray14, [5] reported preliminary localisation results using class activation mappings (CAM) and a static thresholding method to extract bounding boxes from a 50-layer ResNet. These results were later improved by [9] who proposed a patch-based network that trained explicitly using a subset of ChestX-ray14's bounding box annotated images. Although [9]'s results demonstrated a significant improvement over [5]'s, it has been noted by [4] and others that due to the prohibitively time-consuming process of obtaining radiologist-annotated examples, there is great incentive for models that localise well from training only on images with global labels.

Considering this fact and that the best currently-known models for classification on the ChestX-ray14 dataset are ResNet-38-large-meta (R38LM) [8] and CheXNet [6], in this work we focused on models trained on global labels and re-implemented and conducted a quantitative localisation study of the R38LM [8] and CheXNet [6] models for both the 'official' dataset split, and on 5 randomly sampled dataset splits. We also examined the impact of several visualisation methods on localisation quality, and the impact of applying different smoothing methods to these visualisations. Finally, we propose a novel dynamic thresholding method for generating bounding boxes from visualisations, comparing the localisations produced by this method to those produced by the static thresholding approach that is typically used.

II. Methodology

We re-implement and train R38LM [8] and CheXNet [6] as per their original publications on the official ChestX-ray14 dataset split and 5 randomly sampled splits, reporting classification results for both models using area under the receiver operating characteristic curve (AUROC). Then, we use four different visualisation methods, with and without smoothing, employing both our proposed dynamic thresholding scheme and the standard static thresholding scheme to generate bounding boxes. We then use these bounding boxes to evaluate localisation performance. A visual example of the bounding box generation procedure can be seen in Fig. 1.

A. The R38LM and CheXNet deep learning models

R38LM [8] consists of a 38-layer ResNet that takes a 448×448 image as input and creates a 2048-d feature vector that is concatenated with a 3-d feature vector representing the metadata associated with the image (view position, patient's age and sex). A fully connected layer with a sigmoid activation function is used for final classification into 15 classes (14 pathologies + *No finding*). CheXNet [6] instead uses a 121-layer DenseNet with a 14-d fully connected layer and a sigmoid activation function to represent only pathology classes. For both models, all data preprocessing steps (including data augmentation) and training parameters follow those outlined in the original publications, in order to replicate the original studies as faithfully as possible. Binary Cross Entropy (BCE) was used as the loss function for training CheXNet with a mini batch size of 16 and an initial learning rate of 0.001, whereas class-averaged BCE was used for R38LM with a mini batch size of 8 and a learning rate of 0.01. The Adam optimiser ($\beta_1 = 0.9, \beta_2 = 0.999$) was used for both, with training stopping after 100 epochs with no improvement to the average AUROC on the validation set. In addition, after 20 epochs with no improvement, R38LM's learning rate was multiplied by 0.5, whereas CheXNet's learning rate was multiplied by 0.1.

B. Dataset

The NIH's ChestX-ray14 dataset [5] is used in this study due to its ubiquity in the literature and since it is the only publicly available dataset with bounding box annotated pathologies, facilitating a quantitative localisation study. The dataset contains 112,120 chest X-rays of 30,805 patients aged 0 to 95, with 14 labels representing the 14 pathologies shown in TABLE I. Images with no associated labels are implicitly considered as *No Finding*, and multiple labels may be associated with a single image. Additionally, there is a subset of 984 images which have been hand-annotated with bounding boxes by board-certified radiologists for the 8 pathologies shown in TABLE II. In addition, the dataset contains metadata detailing each patient's age, sex and whether the X-ray was taken from front-to-back (anteroposterior) or back-to-front (posteroanterior). For our experiments, we used both the official training/test split, as well as 5 randomly sampled train (70%), validation (10%), test (20%) splits, ensuring that images of the same patient are contained in exactly one of the sets.

C. Visualisation methods

We study the localisation performance of R38LM and CheXNet using the following visualisation methods: class-activation mapping (CAM) [10], Gradient-weighted CAM (Grad-CAM) [11], Grad-CAM++ [12], and Eigen-CAM [13]. Let the final convolutional layer in our network produce a set of activations of size $H \times W \times K$, where $H \times W$ are the dimensions of each grid of activations and K is the total number of such activation grids produced. We denote the k-th grid of activation values produced by this layer as $A^k \in \mathbb{R}^{H \times W}$. For a class c, we denote the weight of the synapse flowing from the k-th neuron in our fully connected layer to the pre-sigmoid classification score y^c for class c as w_k^c. Then, the CAM for class c is $M_{\text{CAM}}^c = \sum_k w_k^c A^k$. Grad-CAM extends this idea, weighting activations by the gradient of the weights flowing into the classification score y^c [11]. Grad-CAM++ instead weights activations by weighted combinations of the positive partial derivatives of y^c [12]. Alternatively, Eigen-CAM applies singular value decomposition to the final convolutional layer's activations, factorising them as $A = USV^T$. The Eigen-CAM is then given as $M_{\text{Eigen-CAM}} = AV_1$, where V_1 is the first eigenvector of V.

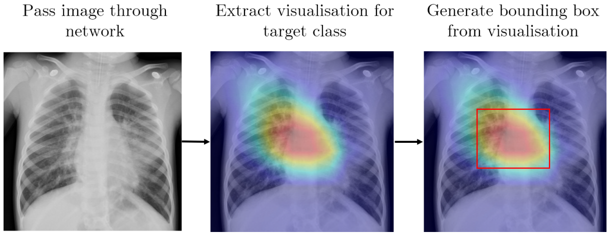

Pass image through network
Extract visualisation for target class
Generate bounding box from visualisation

Fig. 1: Example of bounding box generation

D. Visualisation smoothing

We also investigated the impact of smoothing visualisations on localisation performance, employing two techniques from [14]: (i) augmentation smoothing, and (ii) Eigen smoothing. Augmentation smoothing seeks to better center visualisations around the pertinent image areas. This is done by creating three copies of the input image, horizontally flipping each one, and multiplying their pixel intensities by 0.9, 1.0, and 1.1, respectively. The images are then passed through the network and visualisations are extracted. After applying the inverse process (flipping and division), they are averaged to produce the final visualisation. Eigen smoothing instead seeks to denoise visualisations, allowing tighter bounding box predictions. This is done by applying singular value decomposition to the visualisation and retaining only the principal components. This procedure is identical to that of Eigen-CAM, except that we operate on a single visualisation instead of K activations. In preliminary experiments, we observed a greater gain in localisation performance when applying both smoothing techniques compared to applying either technique independently. As a result, we use both augmentation and Eigen smoothing in our smoothed visualisation experiments.

E. Predicting bounding boxes

We evaluate the accuracy of bounding boxes produced using the typical 'static' thresholding scheme and our proposed 'dynamic' thresholding scheme. Under both schemes, once a visualisation has been extracted, the visualisation's pixel intensities are normalised to the range [0, 255]. The static approach consists of selecting a static range (a common choice is [60, 180], also used by [5]), and setting the pixel intensities that fall within this range to 1, and all others to 0. In our proposed dynamic scheme, a percentile value n is chosen, and

TABLE I: AUROC scores for our experiments, Wang et al.'s [5], and Li et al.'s [9] (higher is better).

Pathology	**R38LM (official)**	**R38LM (random)**	**CheXNet (official)**	**CheXNet (random)**	**Wang [5] (official)**	**Li [9] (official)**
Atelectasis	0.750	**0.767**	0.726	0.761	0.707	0.727
Cardiomegaly	0.845	0.857	0.872	**0.886**	0.810	0.836
Effusion	0.811	**0.845**	0.794	**0.845**	0.759	0.789
Infiltration	0.692	**0.705**	0.680	0.696	0.661	0.672
Mass	0.795	**0.810**	0.788	0.791	0.693	0.776
Nodule	0.726	**0.740**	0.720	0.709	0.669	0.696
Pneumonia	0.706	**0.723**	0.683	0.707	0.658	0.649
Pneumothorax	0.820	**0.840**	0.802	0.830	0.799	0.808
Consolidation	0.724	**0.764**	0.717	0.762	0.703	0.720
Edema	0.822	0.845	0.829	**0.853**	0.805	0.806
Emphysema	0.844	0.871	0.848	0.846	0.833	**0.888**
Fibrosis	0.741	0.756	0.782	0.773	**0.786**	0.771
Pleural thickening	0.709	0.747	0.739	**0.753**	0.684	0.737
Hernia	**0.876**	0.839	0.871	0.843	0.872	0.693
Average	0.776	**0.794**	0.775	0.790	0.746	0.755

Note: the dataset split is given in parentheses. Results for random are averaged over our 5 random splits.

we set pixel intensities to 1 if they fall into the top $(100-n)$-th percent of normalised pixel intensities, and 0 otherwise. For both schemes, after thresholding, a bounding box is drawn around the largest contiguous group of 1-valued pixels.

We introduce dynamic thresholding as a result of preliminary experimentation with static thresholds, where we observed strong localisation performance for R38LM using a threshold range of [60, 180], but poor localisations for CheXNet under the same scheme. We found using a higher threshold range of [180, 255] to improve localisations for CheXNet, but substantially worsen localisations for R38LM. By instead using dynamic thresholding, we allow a fixed percentile of pixel intensities to be retained, which results in more consistent localisation quality when applied to different models, as shown in our experimental evaluation.

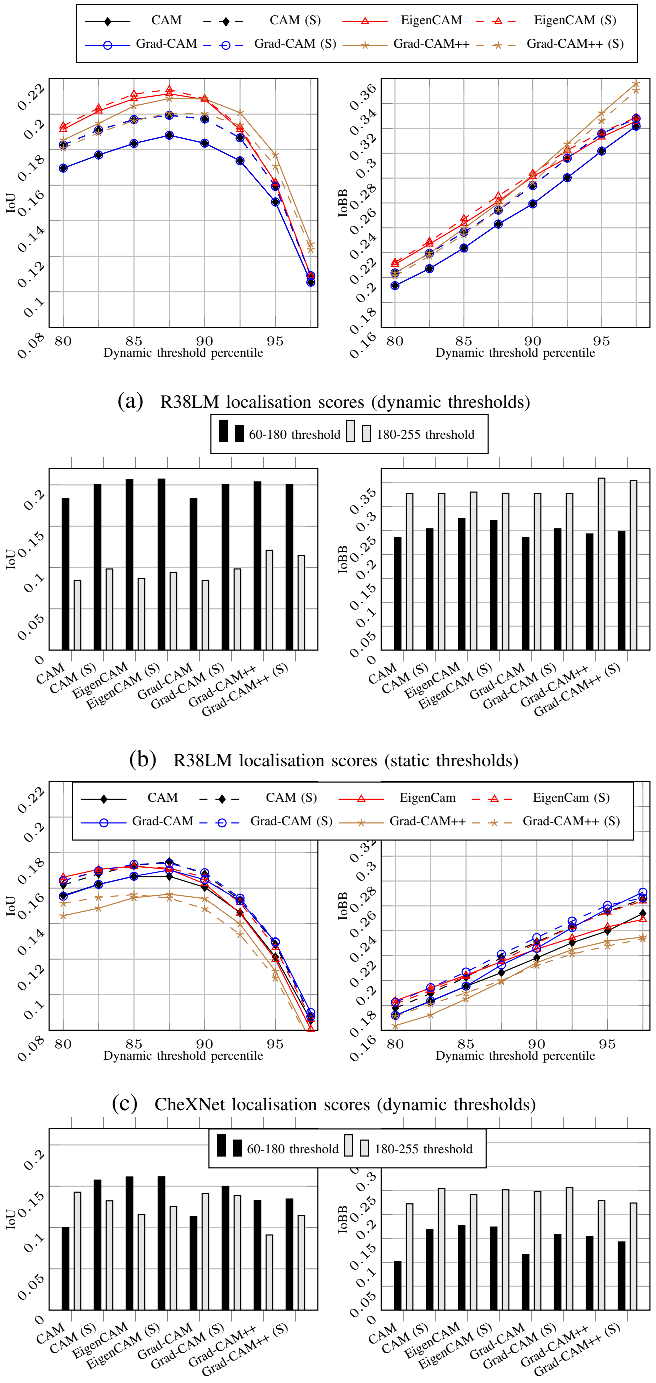
Finally, we measure the quality of predicted bounding boxes using *Intersection over Union* (IoU) and *Intersection over the detected Bounding Box* (IoBB).

III. Results & Discussion

A. Classification results

We report classification performance for R38LM and CheXNet in TABLE I. We obtained results within 5% of those reported by [8] for R38LM, however our CheXNet results were almost 8% worse than those reported by [6], even after repeating our experiments several times on the official split with varying sizes and samples of validation data. As observed by [8], [9] and demonstrated by our AUROC scores on our random splits (which contain proportionally more pathological examples for training than the official split), both R38LM and CheXNet classify better when using a dataset split with more pathological examples in the training set. We expect that our lower than expected CheXNet results are therefore due to a difference in the number of training examples in [5]'s official split and the unspecified random split used in [6]'s original CheXNet experiments.

B. Dynamic vs. static thresholding



(a) R38LM localisation scores (dynamic thresholds)

(b) R38LM localisation scores (static thresholds)

(c) CheXNet localisation scores (dynamic thresholds)

(d) CheXNet localisation scores (static thresholds)

Fig. 2: R38LM's and CheXNet's localisation performance for different visualisation methods under static and dynamic thresholding. (S) refers to visualisation smoothing.

In Fig. 2 we report average localisation scores for the set of bounding box annotated images in ChestX-ray14. For both R38LM and CheXNet, dynamic thresholding with $n \in \{85, 87.5, 90\}$ resulted in higher IoU scores than any static threshold range for all visualisation methods tested. We also observe that when using a sufficiently high value of n, IoBB scores are competitive with or superior to those obtained by static thresholding on both models. Our results demonstrate an inherent limitation of static thresholding, which is that

TABLE II: IoU scores for R38LM and CheXNet on the official split (higher is better).

Pathology	**T(IoU) = 0.1**		**T(IoU) = 0.2**		**T(IoU) = 0.3**		**T(IoU) = 0.4**		**T(IoU) = 0.5**		**T(IoU) = 0.6**		**T(IoU) = 0.7**	
	R38LM	CheXNet	R38LM	CheXNet	R38LM	CheXNet	R38LM	CheXNet	R38LM	CheXNet	R38LM	CheXNet	R38LM	CheXNet
Atelectasis	**0.394**	0.294	**0.194**	0.117	**0.100**	0.050	**0.044**	0.006	**0.028**	0.006	**0.011**	0.006	**0.011**	0.000
Cardiomegaly	**1.000**	0.911	**0.993**	0.911	**0.993**	0.870	**0.938**	0.842	**0.842**	0.712	**0.596**	0.555	**0.281**	0.267
Effusion	**0.536**	0.320	**0.242**	0.144	**0.098**	0.052	**0.039**	0.020	**0.007**	0.000	0.000	0.000	0.000	0.000
Infiltration	**0.862**	0.610	**0.610**	0.455	**0.350**	0.252	**0.179**	0.122	**0.098**	0.065	**0.041**	0.033	**0.008**	**0.008**
Mass	**0.329**	0.282	**0.153**	**0.153**	**0.094**	0.071	**0.047**	0.035	**0.047**	0.012	**0.012**	0.000	**0.012**	0.000
Nodule	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Pneumonia	**0.883**	0.683	**0.625**	0.417	**0.317**	0.258	0.125	**0.142**	0.067	**0.092**	0.025	**0.042**	**0.017**	**0.017**
Pneumothorax	**0.163**	0.133	**0.082**	0.051	**0.031**	**0.031**	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Average	**0.521**	0.404	**0.362**	0.281	**0.248**	0.198	**0.172**	0.146	**0.136**	0.111	**0.086**	0.079	**0.041**	0.036

*Best score per pathology and threshold shown in bold

TABLE III: IoU scores for R38LM and CheXNet on our 5 randomly sampled splits (higher is better).

Pathology	**T(IoU) = 0.1**		**T(IoU) = 0.2**		**T(IoU) = 0.3**		**T(IoU) = 0.4**		**T(IoU) = 0.5**		**T(IoU) = 0.6**		**T(IoU) = 0.7**	
	R38LM	CheXNet	R38LM	CheXNet	R38LM	CheXNet	R38LM	CheXNet	R38LM	CheXNet	R38LM	CheXNet	R38LM	CheXNet
Atelectasis	**0.430**	0.370	**0.224**	0.180	**0.112**	0.082	**0.046**	0.027	**0.026**	0.006	**0.016**	0.003	**0.004**	0.000
Cardiomegaly	**0.999**	0.984	**0.974**	0.967	**0.925**	0.916	**0.858**	0.808	**0.693**	0.679	0.396	**0.489**	0.132	**0.259**
Effusion	**0.571**	0.465	**0.323**	0.235	**0.136**	0.106	**0.044**	0.034	**0.012**	0.007	**0.007**	0.004	**0.001**	0.000
Infiltration	**0.820**	0.722	**0.615**	0.501	**0.350**	0.324	**0.182**	**0.182**	**0.096**	0.076	**0.036**	0.024	**0.015**	0.002
Mass	**0.353**	0.336	0.179	**0.202**	**0.113**	0.111	**0.061**	0.056	0.026	**0.031**	**0.007**	0.002	0.000	0.000
Nodule	**0.013**	0.010	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Pneumonia	**0.862**	0.772	**0.612**	0.500	**0.307**	0.283	**0.158**	0.127	**0.088**	0.062	**0.045**	0.022	**0.015**	0.003
Pneumothorax	**0.192**	0.180	**0.108**	0.084	0.041	**0.049**	**0.022**	0.018	**0.004**	0.000	0.000	0.000	0.000	0.000
Average	**0.530**	0.480	**0.379**	0.334	**0.248**	0.234	**0.171**	0.157	**0.118**	0.108	0.063	**0.068**	0.021	**0.033**

*Best score per pathology and threshold shown in bold

while a specific threshold range may lead to better localisation performance on one model, it can perform poorly on another. For example, the static threshold range [180, 255] yields reasonably similar IoU scores for CheXNet as the threshold range [60, 180], but when using the threshold range [180, 255] on R38LM, the IoU is almost half of that achieved when using [60, 180] instead. As a result, studies that compare multiple models using the same static thresholding approach (e.g. [5]) may be underestimating the true localisation capabilities of these models. Our results also seem to indicate that localisation quality is more sensitive to the thresholding scheme than visualisation method used. While further work is needed to confirm the utility of dynamic thresholding schemes, it appears that dynamic thresholding allows fairer evaluation of localisation performance between models than static thresholding, and has the benefit of more faithfully representing their ability to localise by producing higher localisation scores.

C. Comparison of visualisation methods

Our results for R38LM in Fig. 2 show smoothed Eigen-CAM to perform the best in terms of IoU and IoBB when using dynamic thresholding with $n \in \{80, 82.5, 85, 87.5, 90\}$ and for both static thresholding ranges. Conversely, for CheXNet, while smoothed Eigen-CAM was one of the best performing visualisation methods, it was narrowly beaten by smoothed CAM and smoothed Grad-CAM in terms of IoU and IoBB when using dynamic thresholding with $n \in \{85, 87.5, 90, 92.5, 95, 97.5\}$. Our results also demonstrate that for all visualisation methods tested, with the exception of Grad-CAM++, the application of smoothing reliably improves IoU and IoBB scores. We use smoothed Eigen-CAM and dynamic thresholding with $n = 87.5$ to generate bounding boxes for our final evaluation of localisation performance, since this configuration produced the highest IoU and IoBB scores when averaging across both models.

TABLE IV: IoBB scores for R38LM and CheXNet on the official split (higher is better).

Pathology	**T(IoBB) = 0.1**		**T(IoBB) = 0.25**		**T(IoBB) = 0.5**		**T(IoBB) = 0.75**		**T(IoBB) = 0.9**	
	R38LM	CheXNet	R38LM	CheXNet	R38LM	CheXNet	R38LM	CheXNet	R38LM	CheXNet
Atelectasis	**0.417**	0.311	**0.144**	0.089	**0.033**	0.011	**0.006**	0.000	**0.006**	0.000
Cardiomegaly	**1.000**	0.918	**1.000**	0.911	**0.986**	0.877	**0.589**	0.486	**0.212**	0.199
Effusion	**0.575**	0.346	**0.255**	0.157	**0.039**	0.013	0.000	0.000	0.000	0.000
Infiltration	**0.862**	0.626	**0.520**	0.382	**0.187**	0.154	**0.130**	0.122	**0.073**	**0.073**
Mass	**0.329**	0.294	**0.153**	0.129	**0.071**	0.035	**0.047**	0.024	**0.024**	0.000
Nodule	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Pneumonia	**0.883**	0.708	**0.500**	0.333	**0.167**	**0.167**	**0.117**	0.092	**0.083**	0.067
Pneumothorax	**0.163**	0.153	**0.082**	0.051	0.010	**0.031**	0.000	0.000	0.000	0.000
Average	**0.529**	0.420	**0.332**	0.257	**0.187**	0.161	**0.111**	0.090	**0.050**	0.042

*Best score per pathology and threshold shown in bold

TABLE V: IoBB scores for R38LM and CheXNet on our 5 randomly sampled splits (higher is better).

Pathology	**T(IoBB) = 0.1**		**T(IoBB) = 0.25**		**T(IoBB) = 0.5**		**T(IoBB) = 0.75**		**T(IoBB) = 0.9**	
	R38LM	CheXNet	R38LM	CheXNet	R38LM	CheXNet	R38LM	CheXNet	R38LM	CheXNet
Atelectasis	**0.437**	0.387	**0.170**	0.144	**0.047**	0.023	**0.007**	0.003	0.000	0.000
Cardiomegaly	**1.000**	0.990	**0.993**	0.977	**0.912**	0.875	0.441	**0.489**	0.063	**0.197**
Effusion	**0.600**	0.516	**0.307**	0.229	**0.065**	0.039	**0.014**	0.007	0.000	**0.001**
Infiltration	**0.833**	0.751	**0.524**	0.446	**0.216**	0.193	**0.124**	0.119	0.060	**0.078**
Mass	**0.365**	0.341	0.172	**0.188**	0.071	**0.075**	**0.026**	0.024	**0.016**	0.009
Nodule	**0.013**	0.010	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Pneumonia	**0.870**	0.805	**0.532**	0.447	**0.185**	0.152	**0.115**	0.103	0.070	**0.073**
Pneumothorax	0.218	**0.227**	**0.104**	0.092	**0.027**	**0.027**	**0.002**	0.000	0.000	0.000
Average	**0.542**	0.503	**0.350**	0.315	**0.190**	0.173	0.091	**0.093**	0.026	**0.045**

*Best score per pathology and threshold shown in bold

D. Localisation results

We report our full localisation results on the official split in TABLE II and IV, and on our 5 randomly sampled splits in TABLE III and V. We report results for multiple thresholds to allow comparison with [5] and [9], where $T(IoU) = 0.1$ represents the proportion of predicted bounding boxes with IoU > 0.1. We observe that R38LM obtained higher average IoU and IoBB scores than CheXNet across all metrics and dataset splits, with CheXNet performing better only when $T(IoBB) \in \{0.75, 0.9\}$. There is also a slight improvement across all metrics for models trained on our randomly sampled splits compared to those trained on the official split, which we expect is due to the greater number of pathological examples in the training sets of our random splits compared to the official split. A surprising observation is that localisation scores for *Infiltration* and *Pneumonia* are very high, despite both models

TABLE VI: $T(IoU) = 0.1$ scores for each pathology. Ours and Wang et al.'s results are on the official split, whereas Li et al. train on 80% of the dataset with patient cross-over.

Pathology	R38LM	CheXNet	Wang et al. [5]	Li et al. [9] (anno.)	Li et al. [9] (w/o anno.)
Atelectasis	0.394	0.294	0.689	**0.732**	0.488
Cardiomegaly	**1.000**	0.911	0.938	0.975	0.989
Effusion	0.536	0.320	0.660	**0.865**	0.693
Infiltration	0.862	0.610	0.707	**0.904**	0.842
Mass	0.329	0.282	0.400	**0.657**	0.342
Nodule	0.000	0.000	0.139	**0.537**	0.081
Pneumonia	**0.883**	0.683	0.633	0.451	0.715
Pneumothorax	0.163	0.133	0.378	**0.594**	0.437
Average	0.521	0.404	0.568	**0.714**	0.573

* 'anno.': training on data supplemented with 80% of the provided bounding box annotations. 'w/o': without.

TABLE VII: $T(IoBB) = 0.1$ scores for each pathology. Ours and Wang et al.'s results are on the official split, whereas Li et al. train on the entire dataset.

Pathology	R38LM	CheXNet	Wang et al. [5]	Li et al. [9] (anno.)	Li et al. [9] (w/o anno.)
Atelectasis	0.417	0.311	0.723	**0.757**	0.630
Cardiomegaly	**1.000**	0.918	0.993	0.987	0.888
Effusion	0.575	0.346	0.712	**0.896**	0.783
Infiltration	0.862	0.626	0.789	**0.950**	0.907
Mass	0.329	0.294	0.435	**0.700**	0.696
Nodule	0.000	0.000	0.165	**0.545**	0.292
Pneumonia	**0.883**	0.708	0.750	0.558	0.306
Pneumothorax	0.163	0.153	0.459	**0.632**	0.436
Average	0.529	0.420	0.628	**0.753**	0.617

* 'anno.': training on data supplemented with 80% of the provided bounding box annotations. 'w/o': without.

showing poor classification ability for them. This suggests that while both models may be poor classifiers for these pathologies, they are still able to spatially locate them, and may therefore still be useful in a clinical setting. Conversely, we observe very poor localisations for *Nodule*, which typically has the smallest bounding boxes of any pathology in ChestX-ray14. Since nodule is one of the pathologies most often missed by radiologists [4], there is great incentive to produce DL models capable of localising positive examples well, but it appears that training on global labels alone (as we have) is not enough to produce sufficiently accurate localisations.

We also compare our localisation results to those of [5] and [9] in TABLE VI and VII. We found R38LM to give state-of-the-art results for localising *Cardiomegaly* and *Pneumonia*, but that on average, localisations for both R38LM and CheXNet were worse than those of [5] and [9]. This is particularly surprising, since the AUROC results we report for R38LM and CheXNet are noticably higher than those of the approaches used by [5] and [9] (TABLE I), indicating that a higher AUROC does not necessarily correspond to better localisations. Since [5] followed a similar experimental procedure to our study, training only on global labels with a 50-layer ResNet (ResNet_v2_50), and using a static thresholding approach for bounding box generation, we can only hypothesise that their superior localisation performance is a result of training using a class-weighted loss function. Similarly, [9]'s patch-based network and utilisation of hand-annotated data during training appear to have allowed much stronger localisations than training on global labels alone.

IV. Conclusion

In this work, we presented a quantitative localisation study for two state-of-the-art classifiers on the ChestX-ray14 dataset: ResNet-38-large-meta and CheXNet. Results indicate that a high discriminative ability (measured by AUROC) is not necessarily sufficient for models to produce useful and accurate localisations. We propose a novel method for bounding box generation, dynamic thresholding, and provide evidence that it allows for more faithful representations of different models' abilities to localise compared to static thresholding, and leads to better localisations. In addition, we compared multiple visualisation methods, showing how all but Grad-CAM++ can be reliably improved by the application of augmentation and Eigen smoothing, and evaluated the accuracy of the localisations produced by each method under different thresholding schemes. Finally, we reported R38LM and CheXNet's localisation performance in terms of IoU and IoBB, comparing our results to those of other studies in the literature. For future work, results indicate that while R38LM and CheXNet provide state-of-the-art classification performance, their localisation capability could be potentially improved by supplementing training data with localised examples.

References

[1] C. D. Mathers and D. Loncar, "Projections of global mortality and burden of disease from 2002 to 2030," *PLoS Medicine*, vol. 3, no. 11, p. e442, 2006.

[2] S. I. Kamel, D. C. Levin, L. Parker, and V. M. Rao, "Utilization trends in noncardiac thoracic imaging, 2002-2014," *Journal of the American College of Radiology*, vol. 14, no. 3, pp. 337–342, 2017.

[3] S. Raoof, D. Feigin, A. Sung, S. Raoof, L. Irugulpati, and E. C. Rosenow III, "Interpretation of plain chest roentgenogram," *Chest*, vol. 141, no. 2, pp. 545–558, 2012.

[4] L. Yao, J. Prosky, E. Poblenz, B. Covington, and K. Lyman, "Weakly supervised medical diagnosis and localization from multiple resolutions," *arXiv:1803.07703*, 2018.

[5] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proc. IEEE CVPR*, 2017, pp. 2097–2106.

[6] P. Rajpurkar *et al.*, "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *arXiv:1711.05225*, 2017.

[7] T. Ozturk, M. Talo, E. A. Yildirim, U. B. Baloglu, O. Yildirim, and U. R. Acharya, "Automated detection of COVID-19 cases using deep neural networks with x-ray images," *Computers in Biology and Medicine*, vol. 121, p. 103792, 2020.

[8] I. M. Baltruschat, H. Nickisch, M. Grass, T. Knopp, and A. Saalbach, "Comparison of deep learning approaches for multi-label chest x-ray classification," *Scientific Reports*, vol. 9, no. 1, pp. 1–10, 2019.

[9] Z. Li *et al.*, "Thoracic disease identification and localization with limited supervision," in *Proc. IEEE CVPR*, 2018, pp. 8290–8299.

[10] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE CVPR*, 2016, pp. 2921–2929.

[11] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE ICCV*, 2017, pp. 618–626.

[12] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in *Proc. IEEE WACV*, 2018, pp. 839–847.

[13] M. B. Muhammad and M. Yeasin, "Eigen-CAM: Class activation map using principal components," in *Proc. IJCNN*. IEEE, 2020, pp. 1–7.

[14] J. Gildenblat and contributors, "PyTorch library for CAM methods," https://github.com/jacobgil/pytorch-grad-cam, 2021.