Learning Visual Context by Comparison

Minchul Kim*[1], Jongchan Park*[1], Seil Na[1],
Chang Min Park[2], and Donggeun Yoo†[1]

[1] Lunit Inc., Republic of Korea
{minchul.kim, jcpark, seil.na, dgyoo}@lunit.io
[2] Seoul National University Hospital, Republic of Korea
cmpark.morphius@gmail.com

Abstract. Finding diseases from an X-ray image is an important yet highly challenging task. Current methods for solving this task exploit various characteristics of the chest X-ray image, but one of the most important characteristics is still missing: the necessity of comparison between related regions in an image. In this paper, we present Attend-and-Compare Module (ACM) for capturing the difference between an object of interest and its corresponding context. We show that explicit difference modeling can be very helpful in tasks that require direct comparison between locations from afar. This module can be plugged into existing deep learning models. For evaluation, we apply our module to three chest X-ray recognition tasks and COCO object detection & segmentation tasks and observe consistent improvements across tasks. The code is available at https://github.com/mk-minchul/attend-and-compare.

Keywords: Context Modeling, Attention Mechanism, Chest X-Ray

1 Introduction

Among a variety of medical imaging modalities, chest X-ray is one of the most common and readily available examinations for diagnosing chest diseases. In the US, more than 35 million chest X-rays are taken every year [20]. It is primarily used to screen diseases such as lung cancer, pneumonia, tuberculosis and pneumothorax to detect them at their earliest and most treatable stage. However, the problem lies in the heavy workload of reading chest X-rays. Radiologists usually read tens or hundreds of X-rays every day. Several studies regarding radiologic errors [28, 9] have reported that 20-30% of exams are misdiagnosed. To compensate for this shortcoming, many hospitals equip radiologists with computer-aided diagnosis systems. The recent developments of medical image recognition models have shown potentials for growth in diagnostic accuracy [26].

With the recent presence of large-scale chest X-ray datasets [37, 18, 19, 3], there has been a long line of works that find thoracic diseases from chest X-rays

*The authors have equally contributed. †Donggeun Yoo is the corresponding author.

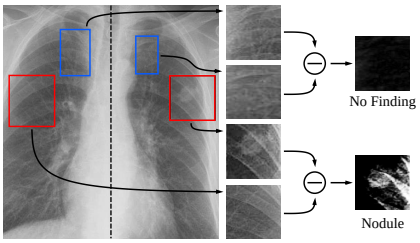


Fig. 1: An example of a comparison procedure for radiologists. Little differences indicate no disease (blue), the significant difference is likely to be a lesion (red).

using deep learning [41, 12, 23, 29]. Most of the works attempt to classify thoracic diseases, and some of the works further localize the lesions. To improve recognition performance, Yao *et al.* [41] handles varying lesion sizes and Mao *et al.* [25] takes the relation between X-rays of the same patient into consideration. Wang *et al.* [35] introduces an attention mechanism to focus on regions of diseases.

While these approaches were motivated by the characteristics of chest X-rays, we paid attention to how radiology residents are trained, which led to the following question: why don't we model the way radiologists read X-rays? When radiologists read chest X-rays, they compare zones [1], paying close attention to any asymmetry between left and right lungs, or any changes between semantically related regions, that are likely to be due to diseases. This comparison process provides contextual clues for the presence of a disease that local texture information may fail to highlight. Fig. 1 illustrates an example of the process. Previous studies [36, 15, 4, 38] proposed various context models, but none addressed the need for the explicit procedure to *compare* regions in an image.

In this paper, we present a novel module, called *Attend-and-Compare Module* (ACM), that extracts features of an object of interest and a corresponding context to explicitly compare them by subtraction, mimicking the way radiologists read X-rays. Although motivated by radiologists' practices, we pose no explicit constraints for symmetry, and ACM learns to compare regions in a data-driven way. ACM is validated over three chest X-ray datasets [37] and object detection & segmentation in COCO dataset [24] with various backbones such as ResNet [14], ResNeXt [40] or DenseNet [16]. Experimental results on chest X-ray datasets and natural image datasets demonstrate that the explicit comparison process by ACM indeed improves the recognition performance.

Contributions To sum up, our major contributions are as follows:

1. We propose a novel context module called ACM that explicitly compares different regions, following the way radiologists read chest X-rays.
2. The proposed ACM captures multiple comparative self-attentions whose difference is beneficial to recognition tasks.
3. We demonstrate the effectiveness of ACM on three chest X-ray datasets [37] and COCO detection & segmentation dataset [24] with various architectures.

2 Related Work

2.1 Context Modeling

Context modeling in deep learning is primarily conducted with the self-attention mechanism [33, 15, 22, 30]. Attention related works are broad and some works do not explicitly pose themselves in the frame of context modeling. However, we include them to highlight different methods that make use of global information, which can be viewed as context.

In the visual recognition domain, recent self-attention mechanisms [15, 34, 7, 22, 38] generate dynamic attention maps for recalibration (e.g., emphasize salient regions or channels). Squeeze-and-Excitation network (SE) [15] learns to model channel-wise attention using the spatially averaged feature. A Style-based Recalibration Module (SRM) [22] further explores the global feature modeling in terms of style recalibration. Convolutional block attention module (CBAM) [38] extends SE module to the spatial dimension by sequentially attending the important location and channel given the feature. The attention values are computed with global or larger receptive fields, and thus, more contextual information can be embedded in the features. However, as the information is aggregated into a single feature by average or similar operations, spatial information from the relationship among multiple locations may be lost.

Works that explicitly tackle the problem of using context stem from using pixel-level pairwise relationships [36, 17, 4]. Such works focus on long-range dependencies and explicitly model the context aggregation from dynamically chosen locations. Non-local neural networks (NL) [36] calculate pixel-level pairwise relationship weights and aggregate (weighted average) the features from all locations according to the weights. The pairwise modeling may represent a more diverse relationship, but it is computationally more expensive. As a result, Global-Context network (GC) [4] challenges the necessity of using all pairwise relationships in NL and suggests to softly aggregate a single distinctive context feature for all locations. Criss-cross attention (CC) [17] for semantic segmentation reduces the computation cost of NL by replacing the pairwise relationship attention maps with criss-cross attention block which considers only horizontal and vertical directions separately. NL and CC explicitly model the pairwise relationship between regions with affinity metrics, but the qualitative results in [36, 17] demonstrate a tendency to aggregate features only among foreground objects or among pixels with similar semantics.

Sharing a similar philosophy, there have been works on contrastive attention [31, 42]. MGCAM [31] uses the contrastive feature between persons and backgrounds, but it requires extra mask supervision for persons. C-MWP [42] is a technique for generating more accurate localization maps in a contrastive manner, but it is not a learning-based method and uses pretrained models.

Inspired by how radiologists diagnose, our proposed module, namely, ACM explicitly models a comparing mechanism. The overview of the module can be found in Fig. 2. Unlike the previous works proposed in the natural image domain, our work stems from the precise need for incorporating difference operation in

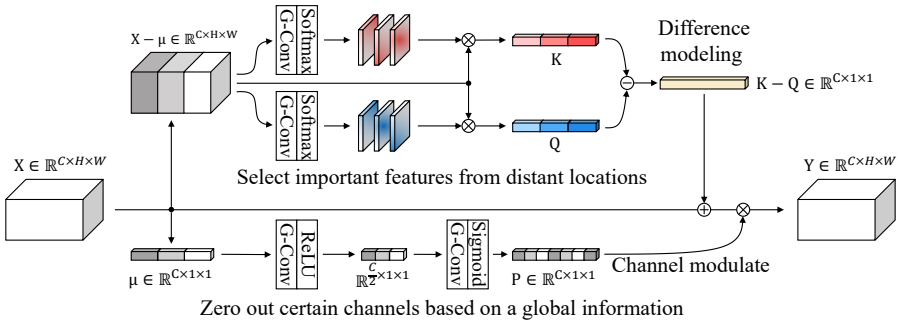


Fig. 2: Illustration of the ACM module. It takes in an input feature and uses the mean-subtracted feature to calculate two feature vectors (K, Q). Each feature vector (K or Q) contains multiple attention vectors from multiple locations calculated using grouped convolutions and normalizations. The difference of the vectors is added to the main feature to make the information more distinguishable. The resulting feature is modulated channel-wise, by the global information feature.

reading chest radiographs. Instead of finding an affinity map based attention as in NL [36], ACM explicitly uses direct comparison procedure for context modeling; instead of using extra supervision to localize regions to compare as in MGCAM [31], ACM automatically learns to focus on meaningful regions to compare. Importantly, the efficacy of our explicit and data-driven contrastive modeling is shown by the superior performance over other context modeling works.

2.2 Chest X-ray as a Context-Dependent Task

Recent releases of the large-scale chest X-ray datasets [37, 18, 19, 3] showed that commonly occurring diseases can be classified and located in a weakly-supervised multi-label classification framework. ResNet [14] and DenseNet [16, 29] pretrained on ImageNet [8] have set a strong baseline for these tasks, and other studies have been conducted on top of them to cover various issues of recognition task in the chest X-ray modality.

To address the issue of localizing diseases using only class-level labels, Guendel *et al.* [12] propose an auxiliary localization task where the ground truth of the location of the diseases is extracted from the text report. Other works use attention module to indirectly align the class-level prediction with the potentially abnormal location [35, 32, 10] without the text reports on the location of the disease. Some works observe that although getting annotations for chest X-rays is costly, it is still helpful to leverage both a small number of location annotations and a large number of class-level labels to improve both localization and classification performances [23]. Guan *et al.* [11] also proposes a hierarchical hard-attention for cascaded inference.

In addition to such characteristics inherent in the chest X-ray image, we would like to point out that the difference between an object of interest and

a corresponding context could be the crucial key for classifying or localizing several diseases as it is important to compare semantically meaningful locations. However, despite the importance of capturing the semantic difference between regions in chest X-ray recognition tasks, no work has dealt with it yet. Our work is, to the best of our knowledge, the first to utilize this characteristic in the Chest X-ray image recognition setting.

3 Attend-and-Compare Module

3.1 Overview

Attend-and-Compare Module (ACM) extracts an object of interest and the corresponding context to compare, and enhances the original image feature with the comparison result. Also, ACM is designed to be light-weight, self-contained, and compatible with popular backbone architectures [14, 16, 40]. We formulate ACM comprising three procedures as

$$Y = f_{\text{ACM}}(X) = P(X + (K - Q)), \tag{1}$$

where f_{ACM} is a transformation mapping an input feature $X \in \mathbb{R}^{C\times H\times W}$ to an output feature $Y \in \mathbb{R}^{C\times H\times W}$ in the same dimension. Between $K \in \mathbb{R}^{C\times 1\times 1}$ and $Q \in \mathbb{R}^{C\times 1\times 1}$, one is intended to be the object of interest and the other is the corresponding context. ACM compares the two by subtracting one from the other, and add the comparison result to the original feature X, followed by an additional channel re-calibration operation with $P \in \mathbb{R}^{C\times 1\times 1}$. Fig. 2 illustrates Equation (1). These three features K, Q and P are conditioned on the input feature X and will be explained in details below.

3.2 Components of ACM

Object of Interest and Corresponding Context To fully express the relationship between different spatial regions of an image, ACM generates two features (K, Q) that focus on two spatial regions of the input feature map X. At first, ACM normalizes the input feature map as $X := X - \mu$ where μ is a C-dimensional mean vector of X. We include this procedure to make training more stable as K and Q will be generated by learnable parameters (W_K, W_Q) that are shared by all input features. Once X is normalized, ACM then calculates K with W_K as

$$K = \sum_{i,j\in H,W} \frac{\exp(W_K X_{i,j})}{\sum_{H,W}\exp(W_K X_{h,w})} X_{i,j}, \tag{2}$$

where $X_{i,j} \in \mathbb{R}^{C\times 1\times 1}$ is a vector at a spatial location (i, j) and $W_K \in \mathbb{R}^{C\times 1\times 1}$ is a weight of a 1×1 convolution. The above operation could be viewed as applying 1×1 convolution on the feature map X to obtain a single-channel attention map in $\mathbb{R}^{1\times H\times W}$, applying softmax to normalize the attention map, and finally weighted averaging the feature map X using the normalized map. Q

is also modeled likewise, but with W_Q. K and Q serve as features representing important regions in X. We add $K - Q$ to the original feature so that the comparative information is more distinguishable in the feature.

Channel Re-calibration In light of the recent success in self-attention modules which use a globally pooled feature to re-calibrate channels [15, 22, 38], we calculate the channel re-calibrating feature P as

$$P = \sigma \circ \text{conv}_2^{1\times1} \circ \text{ReLU} \circ \text{conv}_1^{1\times1}(\mu), \tag{3}$$

where σ and $\text{conv}^{1\times1}$ denote a sigmoid function and a learnable 1×1 convolution function, respectively. The resulting feature vector P will be multiplied to $X + (K - Q)$ to scale down certain channels. P can be viewed as marking which channels to attend with respect to the task at hand.

Group Operation To model a relation of multiple regions from a single module, we choose to incorporate group-wise operation. We replace all convolution operations with grouped convolutions [21, 40], where the input and the output are divided into G number of groups channel-wise, and convolution operations are performed for each group separately. In our work, we use the grouped convolution to deliberately represent multiple important locations from the input. Here, we compute G different attention maps by applying grouped convolution to X, and then obtain the representation $K = [K^1, \cdots, K^G]$ by aggregating each group in X with each attention as follows:

$$K^g = \sum_{i,j \in H,W} \frac{\exp(W_K^g X_{i,j}^g)}{\sum_{H,W} \exp(W_K^g X_{h,w}^g)} X_{i,j}^g, \tag{4}$$

where g refers to g-th group.

Loss Function ACM learns to utilize comparing information within an image by modeling $\{K, Q\}$ whose difference can be important for the given task. To further ensure diversity between them, we introduce an orthogonal loss. based on a dot product. It is defined as

$$\ell_{\text{orth}}(K, Q) = \frac{K \cdot Q}{C}, \tag{5}$$

where C refers to the number of channels. Minimizing this loss can be viewed as decreasing the similarity between K and Q. One trivial solution to minimizing the term would be making K or Q zeros, but they cannot be zeros as they come from the weighted averages of X. The final loss function for a target task can be written as

$$\ell_{\text{task}} + \lambda \sum_{m}^{M} \ell_{\text{orth}}(K_m, Q_m), \tag{6}$$

where ℓ_{task} refers to a loss for the target task, and M refers to the number of ACMs inserted into the network. λ is a constant for controlling the effect of the orthogonal constraint.

Placement of ACMs In order to model contextual information in various levels of feature representation, we insert multiple ACMs into the backbone network. In ResNet, following the placement rule of SE module [15], we insert the module at the end of every Bottleneck block. For example, a total of 16 ACMs are inserted in ResNet-50. Since DenseNet contains more number of DenseBlocks than ResNet's Bottleneck block, we inserted ACM in DenseNet every other three DenseBlocks. Note that we did not optimize the placement location or the number of placement for each task. While we use multiple ACMs, the use of grouped convolution significantly reduces the computation cost in each module.

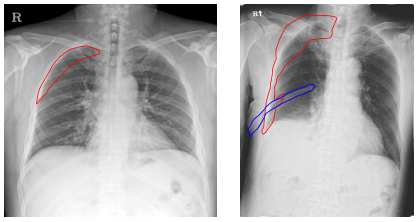
4 Experiments

We evaluate ACM in several datasets: internally-sourced Emergency-Pneumothorax (Em-Ptx) and Nodule (Ndl) datasets for lesion localization in chest X-rays, Chest X-ray14 [37] dataset for multi-label classification, and COCO 2017 [24] dataset for object detection and instance segmentation. The experimental results show that ACM outperforms other context-related modules, in both chest X-ray tasks and natural image tasks.

Experimental Setting Following the previous study [2] on multi-label classification with chest X-Rays, we mainly adopt ResNet-50 as our backbone network. To show generality, we sometimes adopt DenseNet [16] and ResNeXt [40] as backbone networks. In classification tasks, we report class-wise Area Under the Receiver Operating Characteristics (AUC-ROC) for classification performances. For localization tasks, we report the jackknife free-response receiver operating characteristic (JAFROC) [5] for localization performances. JAFROC is a metric widely used for tracking localization performance in radiology. All chest X-ray tasks are a weakly-supervised setting [27] in which the model outputs a probability map for each disease, and final classification scores are computed by global maximum or average pooling. If any segmentation annotation is available, extra map losses are given on the class-wise confidence maps. For all experiments, we initialize the backbone weights with ImageNet-pretrained weights and randomly initialize context-related modules. Experiment details on each dataset are elaborated in the following section.

4.1 Localization on Em-Ptx Dataset

Task Overview The goal of this task is to localize emergency-pneumothorax (Em-Ptx) regions. Pneumothorax is a fatal thoracic disease that needs to be treated immediately. As a treatment, a medical tube is inserted into the pneumothorax affected lung. It is often the case that a treated patient repeatedly takes chest X-rays over a short period to see the progress of the treatment. Therefore, a chest X-ray with pneumothorax is categorized as an emergency, but a chest X-ray with both pneumothorax and a tube is not an emergency. The goal of the task is to correctly classify and localize emergency-pneumothorax.



(a) Emergency (b) Non-emergency

Fig. 3: Examples of pneumothorax cases and annotation maps in Em-Ptx dataset. Lesions are drawn in red. (a) shows a case with pneumothorax, and (b) shows a case which is already treated with a medical tube marked as blue.

To accurately classify the emergency-pneumothorax, the model should exploit the relationship between pneumothorax and tube within an image, even when they are far apart. In this task, utilizing the context as the presence/absence of a tube is the key to accurate classification.

We internally collected 8,223 chest X-rays, including 5,606 pneumothorax cases, of which 3,084 cases are emergency. The dataset is from a real-world cohort and contains cases with other abnormalities even if they do not have pneumothorax. We received annotations for 10 major x-ray findings (Nodule, Consolidation, etc) and the presence of medical devices (EKG, Endotracheal tube, Chemoport, etc). Their labels were not used for training. The task is a binary classification and localization of emergency-pneumothorax. All cases with pneumothorax and tube together and all cases without pneumothorax are considered a non-emergency. We separated 3,223 cases as test data, of which 1,574 cases are emergency-pneumothorax, 1,007 cases are non-emergency-pneumothorax, and 642 cases are pure normal. All of the 1,574 emergency cases in the test data were annotated with coarse segmentation maps by board-certified radiologists. Of the 1,510 emergency-pneumothorax cases in the training data, only 930 cases were annotated, and the rest were used with only the class label. An example of a pneumothorax case and an annotation map created by board-certified radiologists is provided in Fig. 3.

Training Details We use Binary Cross-Entropy loss for both classification and localization, and SGD optimizer with momentum 0.9. The initial learning rate is set to 0.01. The model is trained for 35 epochs in total and the learning rate is dropped by the factor of 10 at epoch 30. For each experiment setting, we report the average AUC-ROC and JAFROC of 5 runs with different initialization.

Result The experimental result is summarized in Table 1. Compared to the baseline ResNet-50 and ResNet-101, all the context modules have shown performance improvements. ACM outperforms all other modules in terms of both AUC-ROC and JAFROC. The result supports our claim that the contextual

Table 1: Results on Em-Ptx dataset. Average of 5 random runs are reported for each setting with standard deviation. RN stands for ResNet [14].

Method	AUC-ROC	JAFROC	Method	AUC-ROC	JAFROC
RN-50	86.78±0.58	81.84±0.64	RN-101	89.75±0.49	85.36±0.44
RN-50 + SE [15]	93.05±3.63	89.19±4.38	RN-101 + SE [15]	90.36±0.83	85.54±0.85
RN-50 + NL [36]	94.63±0.39	91.93±0.68	RN-101 + NL [36]	94.24±0.34	91.70±0.83
RN-50 + CC [17]	87.73±8.66	83.32±10.36	RN-101 + CC [17]	92.57±0.89	89.75±0.89
RN-50 + ACM	**95.35**±0.12	**94.16**±0.21	RN-101 + ACM	**95.43**±0.14	**94.47**±0.10

Table 2: Performance with respect to varying module architectures and hyper-parameters on Em-Ptx dataset. All the experiments are based on ResNet-50.

Module	AUC-ROC	JAFROC
None	86.78±0.58	81.84±0.64
$X + (K - Q)$	94.25±0.31	92.94±0.36
PX	87.16±0.42	82.05±0.30
$P(X + K)$	94.96±0.15	93.59±0.24
$P(X + (K - Q))$	**95.35**±0.12	**94.16**±0.21

(a) Ablations on K, Q and P.

#groups	AUC-ROC	JAFROC
8	90.96±1.88	88.79±2.23
32	**95.35**±0.12	**94.16**±0.21
64	95.08±0.25	93.73±0.31
128	94.89±0.53	92.88±0.53

(b) Ablations on number of groups.

λ	AUC-ROC	JAFROC
0.00	95.11±0.20	93.87±0.20
0.01	95.29±0.34	94.09±0.41
0.10	**95.35**±0.12	**94.16**±0.21
1.00	95.30±0.17	94.04±0.11

(c) Ablations on orthogonal loss weight λ.

information is critical to Em-Ptx task, and our module, with its explicit feature-comparing design, shows the biggest improvement in terms of classification and localization.

4.2 Analysis on ACM with Em-Ptx dataset

In this section, we empirically validate the efficacy of each component in ACM and search for the optimal hyperparameters. For the analysis, we use the Em-Ptx dataset. The training setting is identical to the one used in Sec. 4.1. The average of 5-runs is reported.

Effect of Sub-modules As described in Sec. 3, our module consists of 2 sub-modules: difference modeling and channel modulation. We experiment with the two sub-modules both separately and together. The results are shown in Table 2. Each sub-module brings improvements over the baseline, indicating the context modeling in each sub-module is beneficial to the given task. Combining the sub-modules brings extra improvement, showing the complementary effect of the two sub-modules. The best performance is achieved when all sub-modules are used.

Number of Groups By dividing the features into groups, the module can learn to focus on multiple regions, with only negligible computational or parametric costs. On the other hand, too many groups can result in too few channels per group, which prevents the module from exploiting correlation among many channels. We empirically find the best setting for the number of groups. The results are summarized in Table 2. Note that, training diverges when the number of groups is 1 or 4. The performance improves with the increasing number of groups and saturates after 32. We set the number of groups as 32 across all other experiments, except in DenseNet121 (Sec. 4.4) where we set the number of channel per group to 16 due to channel divisibility.

Orthogonal Loss Weight We introduce a new hyperparameter λ to balance between the target task loss and the orthogonal loss. Although the purpose of the orthogonal loss is to diversify the compared features, an excessive amount of λ can disrupt the trained representations. We empirically determine the magnitude of λ. Results are summarized in Table 2. From the results, we can empirically verify the advantageous effect of the orthogonal loss on the performance, and the optimal value of λ is 0.1. In the validation set, the average absolute similarities between K and Q with $\lambda = 0, 0.1$ are 0.1113 and 0.0394, respectively. It implies that K and Q are dissimilar to some extent, but the orthogonal loss further encourages it. We set λ as 0.1 across all other experiments.

4.3 Localization on Ndl Dataset

Task Overview The goal of this task is to localize lung nodules (Ndl) in chest X-ray images. Lung nodules are composed of fast-growing dense tissues and thus are displayed as tiny opaque regions. Due to inter-patient variability, view-point changes and differences in imaging devices, the model that learns to find nodular patterns with respect to the normal side of the lung from the same image (context) may generalize better. We collected 23,869 X-ray images, of which 3,052 cases are separated for testing purposes. Of the 20,817 training cases, 5,817 cases have nodule(s). Of the 3,052 test cases, 694 cases are with nodules. Images without nodules may or may not contain other lung diseases. All cases with nodule(s) are annotated with coarse segmentation maps by board-certified radiologists. We use the same training procedure for the nodule localization task as for the Em-Ptx localization. We train for 25 epochs with the learning rate dropping once at epoch 20 by the factor of 10.

Results The experimental result is summarized in Table 3. ACM outperforms all other context modeling methods in terms of both AUC-ROC and JAFROC. The results support our claim that the comparing operation, motivated by how radiologists read X-rays, provides a good contextual representation that helps with classifying and localizing lesions in X-rays. Note that the improvements in this dataset may seem smaller than in the Em-Ptx dataset. In the usage of context modules in general, the bigger increase in performance in Em-Ptx

Table 3: Results on Ndl dataset. Average of 5 random runs are reported for each setting with standard deviation.

Method	AUC-ROC	JAFROC
ResNet-50	87.34±0.34	77.35±0.50
ResNet-50 + SE [15]	87.66±0.40	77.57±0.44
ResNet-50 + NL [36]	88.35±0.35	80.51±0.56
ResNet-50 + CC [17]	87.72±0.18	78.63±0.40
ResNet-50 + ACM	**88.60**±0.23	**83.03**±0.24

dataset is because emergency classification requires knowing both the presence of the tube and the presence of Ptx even if they are far apart. So the benefit of the contextual information is directly related to the performance. However, nodule classification can be done to a certain degree without contextual information. Since not all cases need contextual information, the performance gain may be smaller.

4.4 Multi-label Classification on Chest X-ray14

Task Overview In this task, the objective is to identify the presence of 14 diseases in a given chest X-ray image. Chest X-ray14 [37] dataset is the first large-scale dataset on 14 common diseases in chest X-rays. It is used as a benchmark dataset in previous studies [41, 12, 29, 6, 25, 23, 32, 10]. The dataset contains 112,120 images from 30,805 unique patients. Image-level labels are mined from image-attached reports using natural language processing techniques (each image can have multi-labels). We split the dataset into training (70%), validation (10%), and test (20%) sets, following previous works [37, 41].

Training Details As shown in Table 4, previous works on CXR14 dataset vary in loss, input size, etc. We use CheXNet [29] implementation[3] to conduct context module comparisons, and varied with the backbone architecture and the input size to find if context modules work in various settings. We use BCE loss with the SGD optimizer with momentum 0.9 and weight decay 0.0001. Although ChexNet uses the input size of 224, we use the input size of 448 as it shows a better result than 224 with DenseNet121. More training details can be found in the supplementary material.

Results Table 5 shows test set performance of ACM compared with other context modules in multiple backbone architectures. ACM achieves the best performance of 85.39 with ResNet-50 and 85.03 with DenseNet121. We also observe that Non-local (NL) and Cross-Criss Attention (CC) does not perform well in DenseNet architecture, but attains a relatively good performance of 85.08 and 85.11 in ResNet-50. On the other hand, a simpler SE module performs well in

[3] https://github.com/jrzech/reproduce-chexnet

Table 4: Reported performance of previous works on CXR14 dataset. Each work differs in augmentation schemes and some even in the usage of the dataset. We choose CheXNet [29] as the baseline model for adding context modules.

Method	Backbone Arch	Loss Family	Input size	Reported AUC (%)
Wang *et al.* [37]	ResNet-50	CE	1,024	74.5
Yao *et al.* [41]	ResNet+DenseNet	CE	512	76.1
Wang and Xia [35]	ResNet-151	CE	224	78.1
Li *et al.* [23]	ResNet-v2-50	BCE	299	80.6
Guendel *et al.* [12]	DenseNet121	BCE	1,024	80.7
Guan *et al.* [10]	DenseNet121	BCE	224	81.6
ImageGCN [25]	Graph Convnet	CE	224	82.7
CheXNet [29]	DenseNet121	BCE	224	84.1

Table 5: Performance in average AUC of various methods on CXR14 dataset. The numbers in the bracket after model names are the input sizes.

Modules	DenseNet121(448)	ResNet-50(448)
None	(CheXNet [29]) 84.54	84.19
SE [15]	84.95	84.53
NL [36]	84.49	85.08
CC [17]	84.43	85.11
ACM	**85.03**	**85.39**

DenseNet but does poorly in ResNet50. However, ACM shows consistency across all architectures. One of the possible reasons is that it provides a context based on a contrasting operation, thus unique and helpful across different architectures.

4.5 Detection and Segmentation on COCO

Task Overview In this experiment, we validate the efficacy of ACM in the natural image domain. Following the previous studies [36, 17, 4, 38], we use COCO dataset [24] for detection and segmentation tasks. Specifically, we use COCO Detection 2017 dataset, which contains more than 200,000 images and 80 object categories with instance-level segmentation masks. We train both tasks simultaneously using Mask-RCNN architecture [13] in Detectron2 [39].

Training Details Basic training details are identical to the default settings in Detectron2 [39]: learning rate of 0.02 with the batch size of 16. We train for 90,000 iterations, drop the learning rate by 0.1 at iterations 60,000 and 80,000. We use COCO2017-train for training and use COCO2017-val for testing.

Results The experimental results are summarized in Table 6. Although originally developed for chest X-ray tasks, ACM significantly improves the detection

Table 6: Results on COCO dataset. All experiments are based on Mask-RCNN [13].

Method	AP^{bbox}	AP^{bbox}_{50}	AP^{bbox}_{75}	AP^{mask}	AP^{mask}_{50}	AP^{mask}_{75}
ResNet-50	38.59	59.36	42.23	35.24	56.24	37.66
ResNet-50+SE [15]	39.10	60.32	42.59	35.72	57.16	38.20
ResNet-50+NL [36]	39.40	60.60	43.02	35.85	57.63	38.15
ResNet-50+CC [17]	39.82	60.97	42.88	36.05	57.82	38.37
ResNet-50+ACM	**39.94**	**61.58**	**43.30**	**36.40**	**58.40**	**38.63**
ResNet-101	40.77	61.67	44.53	36.86	58.35	39.59
ResNet-101+SE [15]	41.30	62.36	45.26	37.38	59.34	40.00
ResNet-101+NL [36]	41.57	62.75	45.39	37.39	59.50	40.01
ResNet-101+CC [17]	**42.09**	63.21	**45.79**	**37.77**	59.98	**40.29**
ResNet-101+ACM	41.76	**63.38**	45.16	37.68	**60.16**	40.19
ResNeXt-101	43.23	64.42	47.47	39.02	61.10	42.11
ResNeXt-101+SE [15]	43.44	64.91	47.66	39.20	61.92	42.17
ResNeXt-101+NL [36]	43.93	65.44	48.20	39.45	61.99	42.33
ResNeXt-101+CC [17]	43.86	65.28	47.74	39.26	62.06	41.97
ResNeXt-101+ACM	**44.07**	**65.92**	**48.33**	**39.54**	**62.53**	**42.44**

and segmentation performance in the natural image domain as well. In ResNet-50 and ResNeXt-101, ACM outperforms all other modules [15, 4, 36]. The result implies that the comparing operation is not only crucial for X-ray images but is also generally helpful for parsing scene information in the natural image domain.

4.6 Qualitative Results

To analyze ACM, we visualize the attention maps for objects of interest and the corresponding context. The network learns to attend different regions, in such a way to maximize the performance of the given task. We visualize the attention maps to see if maximizing performance is aligned with producing attention maps that highlight interpretable locations. Ground-truth annotation contours are also visualized.

We use the Em-Ptx dataset and COCO dataset for the analysis. Since there are many attention maps to check, we sort the maps by the amount of overlap between each attention map and the ground truth location of lesions. We visualize the attention map with the most overlap. Qualitative results of other tasks are included in the supplementary material due to a space limit.

Pneumothorax is a collapsed lung, and on the X-ray image, it is often portrayed as a slightly darker region than the normal side of the lung. A simple way to detect pneumothorax is to find a region slightly darker than the normal side. ACM learns to utilize pneumothorax regions as objects of interest and normal lung regions as the corresponding context. The attention maps are visualized in Fig. 4. It clearly demonstrates that the module attends to both pneumothorax regions and normal lung regions and compare the two sets of regions. The observation is coherent with our intuition that comparing can help recognize, and indicates that ACM automatically learns to do so.

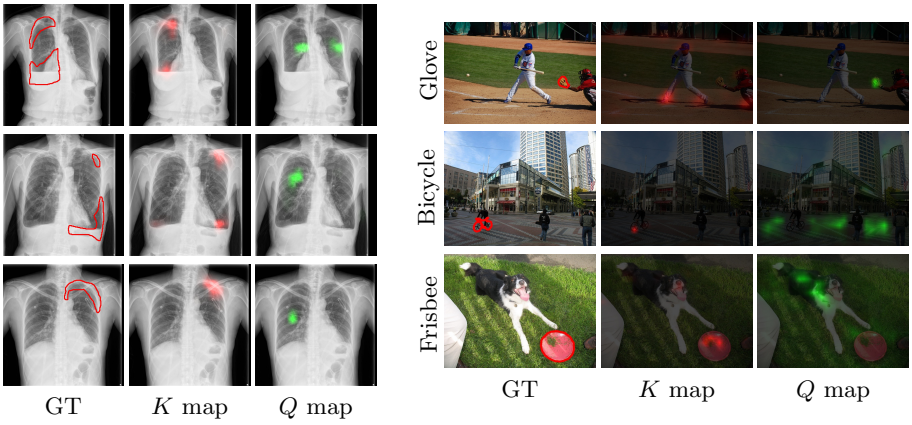


Fig. 4: Left: The visualized attention maps for the localization task on Em-Ptx dataset. The 11th group in the 16th module is chosen. Em-Ptx annotations are shown as red contours on the chest X-ray image. Right: The visualization on COCO dataset. Ground-truth segmentation annotations for each category are shown as red contours.

We also visualize the attended regions in the COCO dataset. Examples in Fig. 4 shows that ACM also learns to utilize the object of interest and the corresponding context in the natural image domain; for *baseball glove*, ACM combines the corresponding context information from the players' heads and feet; for *bicycle*, ACM combines information from roads; for *frisbee*, ACM combines information from dogs. We observe that the relationship between the object of interest and the corresponding context is mainly from co-occurring semantics, rather than simply visually similar regions. The learned relationship is aligned well with the design principle of ACM; selecting features whose semantics differ, yet whose relationship can serve as meaningful information.

5 Conclusion

We have proposed a novel self-contained module, named Attend-and-Compare Module (ACM), whose key idea is to extract an object of interest and a corresponding context and explicitly compare them to make the image representation more distinguishable. We have empirically validated that ACM indeed improves the performance of visual recognition tasks in chest X-ray and natural image domains. Specifically, a simple addition of ACM provides consistent improvements over baselines in COCO as well as Chest X-ray14 public dataset and internally collected Em-Ptx and Ndl dataset. The qualitative analysis shows that ACM automatically learns dynamic relationships. The objects of interest and corresponding contexts are different yet contain useful information for the given task.The qualitative analysis shows that ACM automatically learns dynamic relationships. The objects of interest and corresponding contexts are different yet contain useful information for the given task.

References

1. Armato III, S.G., Giger, M.L., MacMahon, H.: Computerized detection of abnormal asymmetry in digital chest radiographs. Medical physics **21**(11), 1761–1768 (1994)
2. Baltruschat, I.M., Nickisch, H., Grass, M., Knopp, T., Saalbach, A.: Comparison of deep learning approaches for multi-label chest x-ray classification. Scientific reports **9**(1), 6381 (2019)
3. Bustos, A., Pertusa, A., Salinas, J.M., de la Iglesia-Vayá, M.: Padchest: A large chest x-ray image dataset with multi-label annotated reports. arXiv (2019)
4. Cao, Y., Xu, J., Lin, S., Wei, F., Hu, H.: Gcnet: Non-local networks meet squeeze-excitation networks and beyond. ICCV (2019)
5. Chakraborty, D.P.: Recent advances in observer performance methodology: jackknife free-response roc (jafroc). Radiation protection dosimetry **114**(1-3), 26–31 (2005)
6. Chen, B., Li, J., Guo, X., Lu, G.: Dualchexnet: dual asymmetric feature learning for thoracic disease classification in chest x-rays. Biomedical Signal Processing and Control **53**, 101554 (2019)
7. Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., Chua, T.S.: Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In: CVPR (2017)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR (2009)
9. Forrest, J.V., Friedman, P.J.: Radiologic errors in patients with lung cancer. Western Journal of Medicine **134**(6), 485 (1981)
10. Guan, Q., Huang, Y.: Multi-label chest x-ray image classification via category-wise residual attention learning. Pattern Recognition Letters (2018)
11. Guan, Q., Huang, Y., Zhong, Z., Zheng, Z., Zheng, L., Yang, Y.: Thorax disease classification with attention guided convolutional neural network. Pattern Recognition Letters **131**, 38 – 45 (2020). https://doi.org/https://doi.org/10.1016/j.patrec.2019.11.040, http://www.sciencedirect.com/science/article/pii/S0167865519303617
12. Guendel, S., Grbic, S., Georgescu, B., Liu, S., Maier, A., Comaniciu, D.: Learning to recognize abnormalities in chest x-rays with location-aware dense networks. In: Iberoamerican Congress on Pattern Recognition. pp. 757–765. Springer (2018)
13. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: ICCV (2017)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
15. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR (2018)
16. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: CVPR (2017)
17. Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., Liu, W.: Ccnet: Criss-cross attention for semantic segmentation. In: ICCV (2019)
18. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: AAAI (2019)
19. Johnson, A.E., Pollard, T.J., Berkowitz, S., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data (2019)

20. Kamel, S.I., Levin, D.C., Parker, L., Rao, V.M.: Utilization trends in noncardiac thoracic imaging, 2002-2014. Journal of the American College of Radiology **14**(3), 337–342 (2017)
21. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS (2012)
22. Lee, H., Kim, H.E., Nam, H.: Srm: A style-based recalibration module for convolutional neural networks. In: ICCV (2019)
23. Li, Z., Wang, C., Han, M., Xue, Y., Wei, W., Li, L.J., Fei-Fei, L.: Thoracic disease identification and localization with limited supervision. In: CVPR (2018)
24. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
25. Mao, C., Yao, L., Luo, Y.: Imagegcn: Multi-relational image graph convolutional networks for disease identification with chest x-rays. arXiv (2019)
26. Nam, J.G., Park, S., Hwang, E.J., Lee, J.H., Jin, K.N., Lim, K.Y., Vu, T.H., Sohn, J.H., Hwang, S., Goo, J.M., Park, C.M.: Development and validation of deep learning–based automatic detection algorithm for malignant pulmonary nodules on chest radiographs. Radiology **290**(1), 218–228 (2019). https://doi.org/10.1148/radiol.2018180237, https://doi.org/10.1148/radiol.2018180237, pMID: 30251934
27. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Is object localization for free?-weakly-supervised learning with convolutional neural networks. In: CVPR (2015)
28. Quekel, L.G., Kessels, A.G., Goei, R., van Engelshoven, J.M.: Miss rate of lung cancer on the chest radiograph in clinical practice. Chest **115**(3), 720–724 (1999)
29. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., et al.: Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv (2017)
30. Roy, A.G., Navab, N., Wachinger, C.: Concurrent spatial and channel 'squeeze & excitation'in fully convolutional networks. In: MICCAI (2018)
31. Song, C., Huang, Y., Ouyang, W., Wang, L.: Mask-guided contrastive attention model for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1179–1188 (2018)
32. Tang, Y., Wang, X., Harrison, A.P., Lu, L., Xiao, J., Summers, R.M.: Attention-guided curriculum learning for weakly supervised classification and localization of thoracic diseases on chest radiographs. In: International Workshop on Machine Learning in Medical Imaging. Springer (2018)
33. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NIPS (2017)
34. Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X.: Residual attention network for image classification. In: CVPR (2017)
35. Wang, H., Xia, Y.: Chestnet: A deep neural network for classification of thoracic diseases on chest radiography. arXiv (2018)
36. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: CVPR (2018)
37. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: CVPR (2017)
38. Woo, S., Park, J., Lee, J.Y., So Kweon, I.: Cbam: Convolutional block attention module. In: ECCV (2018)
39. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. https://github.com/facebookresearch/detectron2 (2019)

40. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: CVPR (2017)
41. Yao, L., Prosky, J., Poblenz, E., Covington, B., Lyman, K.: Weakly supervised medical diagnosis and localization from multiple resolutions. arXiv (2018)
42. Zhang, J., Bargal, S.A., Lin, Z., Brandt, J., Shen, X., Sclaroff, S.: Top-down neural attention by excitation backprop. International Journal of Computer Vision **126**(10), 1084–1102 (2018)