Learning Visual Context by Comparison Supplementary Material

In this supplementary material, we include qualitative results and detailed experiment statistics which are not included in the main paper due to the space limit. Qualitative results for Ndl, CXR14, COCO datasets are included in Section A, B, C, respectively. The full table for AUC-ROC in the CXR14 dataset is included in Section B.2.

A Qualitative Results on Ndl dataset

A.1 ACM Attention Map Visualizations

Nodules in chest X-ray images are displayed as discrete, marginated and rounded regions with high opacity. However, the overall opacity of the chest X-ray images can be affected by the type of devices, patients' view positions, or normalization ranges (i.e. window range). The samples in Figure 1 show the attention maps in the 15th module and the 15th group. The attention maps for K focus on the nodule area, and the ones for Q focus on the upper lung regions. Upper lung regions have overlapping bones, so their opacity is consistently higher than in other regions in the lung. The samples indicate that ACM learns to compare suspicious nodule regions and upper lung regions, to figure out characteristics of nodules other than the high opacity.

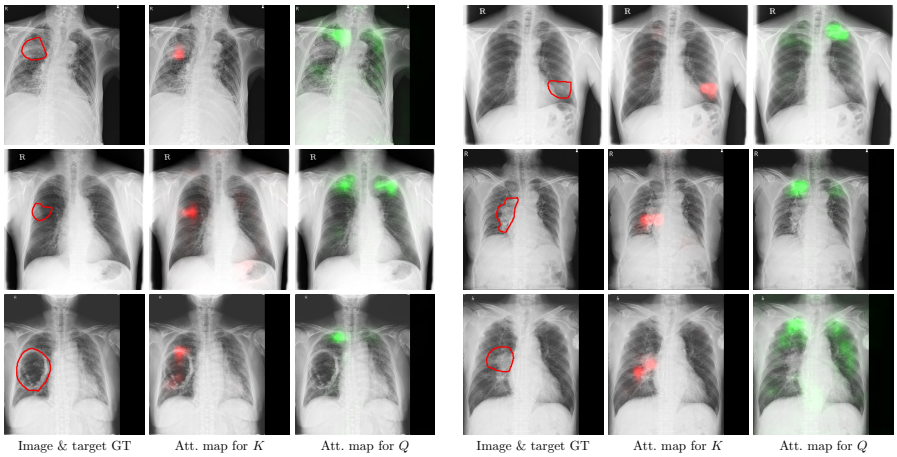


Fig. 1: The visualized attention maps for the localization task in Ndl dataset. Ground-truth segmentation annotations for each category are shown as red contours.

A.2 ACM as a Correction for Failed Prediction

We also analyze samples that ACM helps to reduce false positives across multiple runs. Figure 2 shows the prediction result of the baseline method (ResNet-50) and ACM for several samples of the Ndl dataset and the attention map for K, Q. In Figure 2, in all cases, the baseline method incorrectly predicts that an opaque region inside the lung is a nodule (i.e. false positives). Those regions are confusing as they contain structures that could be mistaken for a nodule. In the 3rd row of Figure 2, the ResNet-50 model predicts that a heart region contains a nodule, a common false-positive observed in chest X-ray modality. On the other hand, ACM captures such a region with an attention map for K, captures another opaque region with an attention map for Q, and compares the two regions to make the final prediction to be suppressed in the confusing region.

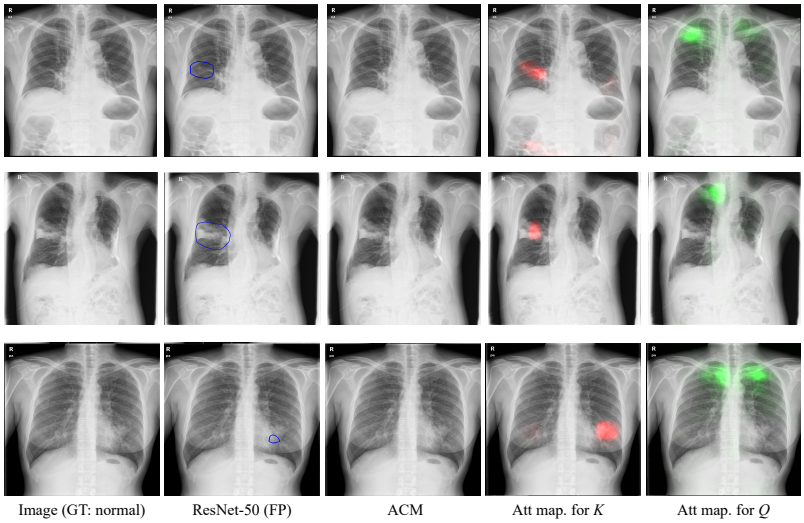


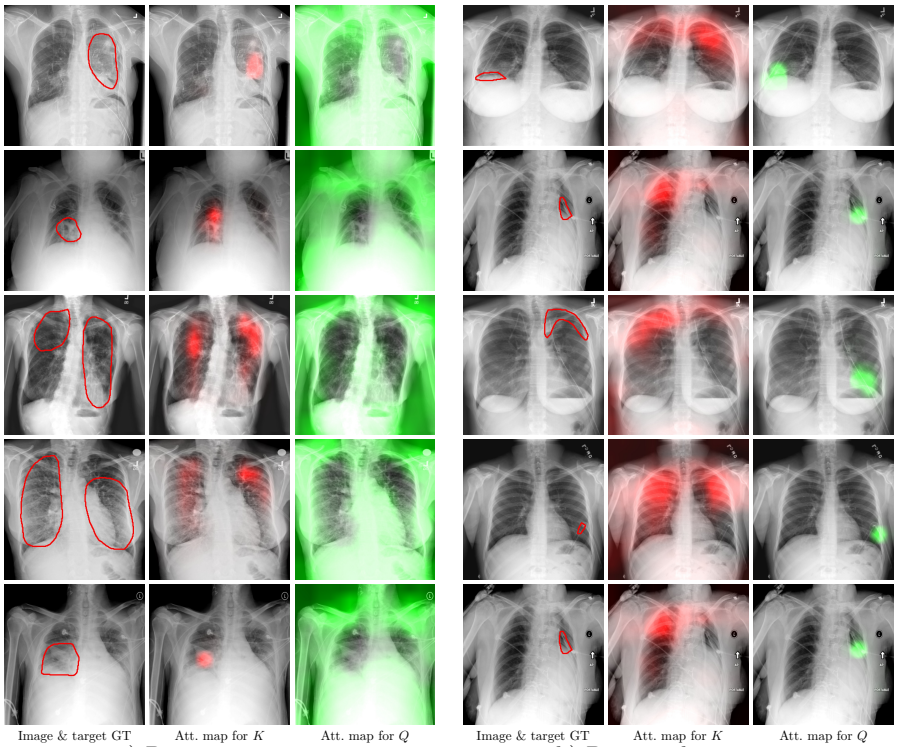
Fig. 2: The visualized localization and attention maps for the failure samples by baseline method (ResNet-50) in localization task in Ndl dataset. FP stands for false positives. Localization predictions by each model are shown as blue contours. The input images do not have nodules (i.e. normal cases in terms of nodules). No indication of a nodule is the correct prediction.

B Qualitative Results on CXR14 dataset and Training Details

B.1 ACM Attention Map Visualizations for Pneumonia and Pneumothorax

CXR14 dataset comes with class-level labels and does not have annotations that are essential for our qualitative analysis. So we randomly selected 500 cases, and choose pneumonia and pneumothorax to be annotated by board-certified radiologists. Pneumonia and pneumothorax are lesions with the most improvement compared to other benchmark methods, according to Table 1. Figure 3 shows samples with pneumonia or pneumothorax. Note that the corresponding context regions are more dispersed that Em-Ptx or Ndl results, which may be due to the weakly-supervised training scheme in the CXR14 dataset.

For pneumonia cases, we show the attention maps from the 15th module and 26th group; For pneumothorax cases, we show the attention maps from the 16th module and 17th group. From Figure 3, we observe that the corresponding context regions for pneumonia are outside the lung. ACM learns to compare suspicious pneumonia regions with outside-lung regions. From Figure 3, we observe that the corresponding context regions for pneumothorax are normal lung regions, mostly upper regions. The observation is similar to the observation in Em-Ptx dataset. However, in some cases, attention maps focus on tubes rather than the pneumothorax, as the common treatment for pneumothorax is a medical tube insertion. It is a bias due to the high spatial correlation between tubes and pneumothorax regions, and in the CXR14 dataset, there are no labels for the tube to compensate for this.



a) Pneumonia cases b) Pneumothorax cases

Fig. 3: The visualized attention maps for CXR14 cases with pneumonia and pneumothorax.

B.2 Full Experiment Table for CXR14

Method	Avg	Atl	Car	Csn	Edm	Eff	Emp	Fib	Hrn	Inf	Ms	Ndl	Pt	Pnm	Ptx
Wang *et al.* [7]	74.5	70.0	81.0	70.3	80.5	75.9	83.3	78.6	87.2	66.1	69.3	66.9	68.4	65.8	79.9
Yao *et al.*. [8]	76.1	73.3	85.6	70.3	80.6	80.6	84.2	74.3	77.5	67.3	77.7	71.8	72.4	68.4	80.5
Wang and Xia [6]	78.1	74.3	87.5	72.3	83.3	81.1	82.2	80.4	90.0	67.7	78.3	69.8	75.1	69.6	81.0
Li *et al.*. [3]	80.6	80.0	87.0	80.0	88.0	87.0	91.0	78.0	77.0	70.0	83.0	75.0	79.0	67.0	87.0
Guendel *et al.*. [2]	80.7	76.7	88.3	74.5	83.5	82.8	89.5	81.8	89.6	70.9	82.1	75.8	76.1	73.1	84.6
Guan *et al.*. [1]	81.6	78.1	88.0	75.4	85.0	82.9	90.8	83.0	91.7	70.2	83.4	77.3	77.8	72.9	85.7
ImageGCN [4]	82.7	79.6	89.6	78.8	88.9	87.3	90.7	81.3	91.7	69.9	83.4	76.2	79.2	71.7	89.0
CheXNet [5]	84.1	80.9	92.5	79.0	88.8	86.4	93.7	80.5	91.6	73.5	86.7	78.0	80.6	76.8	88.9
ACM	**85.4**	83.4	90.7	80.1	90.2	88.7	94.8	85.1	94.8	71.9	86.3	81.5	80.1	77.3	89.8

Table 1: Detailed performance comparison with previous works for chest X-ray 14 dataset. Benchmark models include previous works that tackled this work. Abbrs: Atl: Atelectasis; Car: Cardiomegaly; Csn: Consolidation; Edm: Edema; Eff: Effusion; Emp: Emphysema; Fib: Fibrosis; Hrn: Hernia; Inf: Infiltration; Ms: Mass; Ndl: Nodule; PT: Pleural Thickening; Pnm: Pneumonia; Ptx:Pneumothorax.

B.3 Training Details for CXR14

Our implementation is based on one of the available source code[1] which reproduces the CheXNet [5] result. This means applying the following data augmentation operations; horizontally flipping, random crop and color jittering on input images. We use the BCE loss and the SGD optimizer with momentum 0.9 and weight decay 0.0001. We decay the learning rate by 0.1 when there is no improvement in validation accuracy for 4 epochs, and stop training when there is no improvement in validation accuracy for 5 epochs. Finally, the test accuracy is measured with the checkpoint with the best validation accuracy.

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

C Qualitative Results on COCO dataset

C.1 ACM Attention Map Visualizations

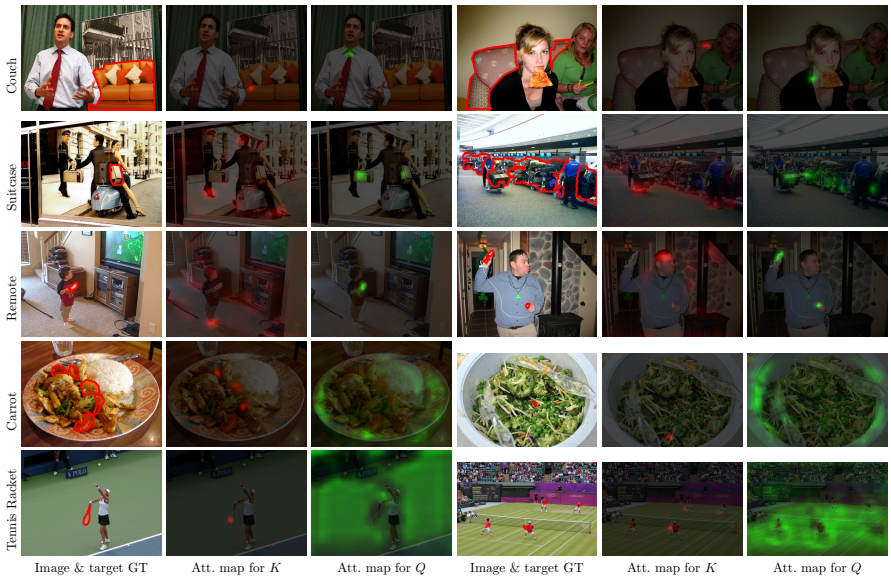


Fig. 4: The visualized attention maps for object detection & segmentation task on COCO dataset.

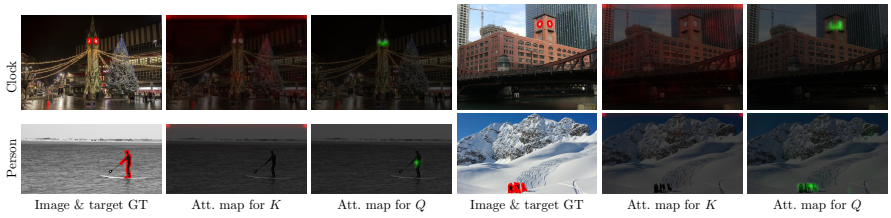


Fig. 5: Samples for classes that does not find corresponding context on COCO dataset.

In addition to the COCO visualization presented in the main paper, we present more visualization results. Many classes have the corresponding context regions that are semantically meaningful and are co-occurring. In Figure 4, we show 2 samples from 5 different classes; *Couch*, *Suitcase*, *Remote*, *Carrot*, and *Tennis Racket*.

For the *Couch*, the module learns to compare the couch-like region with people. The *Couch* class aggregates the corresponding context mostly around the neck region. For the *Suitcase* class, the module can compare the suitcase region with a person's feet, face or wheels of the carts that carry the suitcases. The *Remote* class aggregates from a person's face and feet.

Many classes use the corresponding context from the region in a human, as the human is one of the most commonly appearing objects in the COCO dataset, and many objects are interpreted in relation to the person who is using them. The *Carrot* class, on the other hand, uses the corresponding context from the plates and cutlery, because they appear often with the plates and cutlery. The *Tennis Racket* class learns to gather the corresponding context from the field area, as tennis rackets appear often on the flat green field background.

However, not all classes seem to have corresponding context, and some samples are shown in Figure 5. We show 2 classes: *Clock* and *Person*. Those classes can easily be recognized by themselves, and may not require the corresponding context. From the visualized attention maps, we observe that the corresponding context appears at the corners. ACM is trained to aggregate the corresponding context with a softmax attention map, so when there is no need for the corresponding context, it learns to avoid aggregating from specific semantics by focusing on the corners.

C.2 COCO Dataset Output Visualizations

We include some output results from the COCO dataset to show the benefits of using ACM. We compare the ResNet-50 model with ResNet-50+ACM model and show the output image along with the ACM attention maps selected according to the most overlapping region with the ground truth. The red contours in the first column show the object of interest in this visualization. The corresponding text labels are shown on the left as the rotated text. The first *Carrot* example is an example of a carrot that is hard to detect because it is out-focused. However, ACM finds it successfully when in comparison with the plate nearby. The *Couch* example on the third row shows an occluded couch which can be difficult to find if not taken in the context of the whole image. ACM finds it successfully by utilizing nearby humans as the corresponding context. Also, ResNet-50+ACM model exceeds other models in segmentation, more so than in the detection task. The Tennis-racket example shows that ACM learns to successfully segment even a hard example.

Fig. 6: Samples for COCO Dataset Outputs.

References

1. Guan, Q., Huang, Y.: Multi-label chest x-ray image classification via category-wise residual attention learning. Pattern Recognition Letters (2018)
2. 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)
3. 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)
4. Mao, C., Yao, L., Luo, Y.: Imagegcn: Multi-relational image graph convolutional networks for disease identification with chest x-rays. arXiv (2019)
5. 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)
6. Wang, H., Xia, Y.: Chestnet: A deep neural network for classification of thoracic diseases on chest radiography. arXiv (2018)
7. 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)
8. Yao, L., Prosky, J., Poblenz, E., Covington, B., Lyman, K.: Weakly supervised medical diagnosis and localization from multiple resolutions. arXiv (2018)