CLASS-BASED ATTENTION MECHANISM FOR CHEST RADIOGRAPH MULTI-LABEL CATEGORIZATION

David Sriker* *Hayit Greenspan** *Jacob Goldberger*†

* Department of Biomedical Engineering, Tel Aviv University, Tel Aviv, Israel
† Faculty of Engineering, Bar-Ilan University, Ramat-Gan, Israel

ABSTRACT

This work focuses on a new methodology for class-based attention, which is an extension to the more common image-based attention mechanism. The class-based attention mechanism learns a different attention mask for each class. This enables to simultaneously apply a different localization procedure for different pathologies in the same image, thus important for a multilabel categorization. We apply the method to detect and localize a set of pathologies in chest Radiographs. The proposed network architecture was evaluated on publicly available X-ray datasets and yielded improved classification results compared to standard image based attention.

***Index Terms*—** X-ray, chest, attention mechanism ,localization

1. INTRODUCTION

Chest Radiographs are the most commonly performed radiological exam in the world with industrialized countries reporting an average 238 erect-view chest X-ray images acquired per 1000 of population annually [1]. Chest radiography is critical for screening, diagnosis, and management of many life threatening diseases. Automated interpretation of the images, at the level of practicing radiologists, could provide a substantial benefit in many medical settings, from improved workflow prioritization and clinical decision support to large-scale screening and global population health initiatives.

Among recent advanced deep learning techniques, the attention mechanism has shown to be a powerful tool and been an essential component of neural architectures in a large number of applications in image classification and segmentation (see e.g. [2]) as well as in natural language processing (see e.g. [3]). The basic idea of attention mechanism comes from the visual attention in the human vision system, where human vision always focuses on selective parts of the visual screen. Most studies combining deep learning and visual attention mechanisms have concentrated on the use of masks. The principle of attention mask is that the key areas in the image data are identified by another layer with a new weight. Following the learning phase the network can identify the areas where attention needs to be paid in each new image, thereby forming attention. Jetley et al. [2] proposed a trainable module for generating attention maps to identify the effective spatial support of the visual information used by CNNs in making their classification decisions. Following the growing interest in exploring attention mechanisms in natural scenes, there are also works that adopted attention to medical images analysis and primarily to image segmentation and classification, see e.g. [4, 5, 6, 7, 8, 9]. An updated review on the attention mechanism in deep learning can be found in [10].

In scenery images there is no prior information on the location of different object classes. A cat can a-priori appear in the image in the same places that a dog can appear. Hence, attention can only be based on the input image and not on the class. In medical images the setting is different: organs are located in a known relative positioning within the body; and many pathologies are well defined via their relative positioning within the organs. In chest radiographs, many pathologies can manifest in specific areas (e.g. heart related findings, aorta related and more).

In this study we introduce a CNN architecture that includes a class based attention mechanism. For each class we allocate a different attention component that specializes in detection of the existence of that class in the image. This proposed class based attention method was evaluated on the publicly available ChestX-ray14 dataset [11] and CheXpert dataset [12] and yielded improved classification results compared to a network based on the standard attention methods.

2. CLASS BASED ATTENTION

CNN is the standard network architecture for image analysis. When attention is incorporated into CNN models, the intermediate feature maps are expected to be used more efficiently while avoiding the need for the convolution layers to solve localization and classification tasks separately. The attention module teaches the network to pay attention to areas which will be hopefully relevant to anatomy, and ignore background areas. Attention based networks perform adaptive feature pooling, where model predictions are conditioned solely on a subset of selected image regions. Attention layer

This research was supported by the Ministry of Science & Technology, Israel.

automatically learns to focus on target structures without additional supervision. At test time, these gates generate soft region proposals implicitly on-the-fly and highlight salient image areas that are useful for a specific task. CNN models with attention can be trained from scratch similarly to the training of standard convolutional network models.

We follow here the attention gate concept that was proposed for natural image classification [2] and was later used [5] for ultrasound scan plane detection. Each attention block receives activation maps' local features, $l_s(x)$ from different convolutional layers $s \in \{1,2,3\}$ of the network as well as the global features, $g(x)$, from the final residual block of the network after projecting the features to match the local ones. We use the term layer here for an entire residual block of convolutional layers. Then, we compute the attention score matrix, c_s, from l_s and g as follows:

$$c_{si}(x) = A_s \cdot (B_s \cdot l_{si}(x) + g(x)) \tag{1}$$

such that x is the raw input image and A_s and B_s are the parameters of the attention mask at layer s that are jointly trained with all other network parameters. The index i represents a 'pixel' at layer s. The attention score (1) can be seen to emphasize the image patches that are highly correlated with the global image representation g that is used for classification. We next compute a soft spatial mask by normalizing the scores using the softmax operation:

$$w_{si}(x) = \frac{\exp(c_{si}(x))}{\sum_j \exp(c_{sj}(x))} \tag{2}$$

where i and j represent the features at the corresponding layer. Given the mask w_s, we compute the weighted average of the local features l_s to generate localized global features g_s:

$$g_s(x) = E_{w_s}(l_s(x)) = \sum_i w_{si}(x) l_{si}(x). \tag{3}$$

Finally, we concatenate the obtained g_s's to form a new feature vector and use it for classification instead of g. A scheme of this image based attention model is shown in Fig. 1(top).

In our approach we want the attention decision to rely on the class in addition to the image. Class based attention is not relevant to tasks such as liver lesion detection or breast lesion detection where a lesion can be found anywhere in the organ and (in today's medicine at least) the location does not indicate whether it is benign or malignant. It is relevant to tasks where the decision process involves detection of a pathology which is expected to be in a fixed location in the organ. The first network layers of a CNN perform low-level processing that does not vary much between tasks applied to similar data types. As the data processing progresses along the network layers, the network is more focused on the detection of different classes. In our model, we keep the first two attention blocks as in [2]. Assuming there are k classes, we replicate the third attention block k times with independent sets of attention parameters to generate k feature vectors. We thus replace Eq. (1) in the third attention layer by a class based layer:

$$c_{3ji}(x) = A_{3j} \cdot (B_{3j} \cdot l_{3i}(x) + g(x)), \qquad j = 1, ..., k \tag{4}$$

such that A_{3j} and B_{3j} are the parameters of the third attentions layer for the case of class j. Given the class based score matrix c_{3j} we compute a class based distribution mask w_{3j} and a class based representation g_{3j} using Eqs. (2) and (3). Finally, a binary classification for class j is done using the concatenated representation (g_1, g_2, g_{3j}). Schemes of CNNs with image based attention and class based attention appear in Fig. 1.

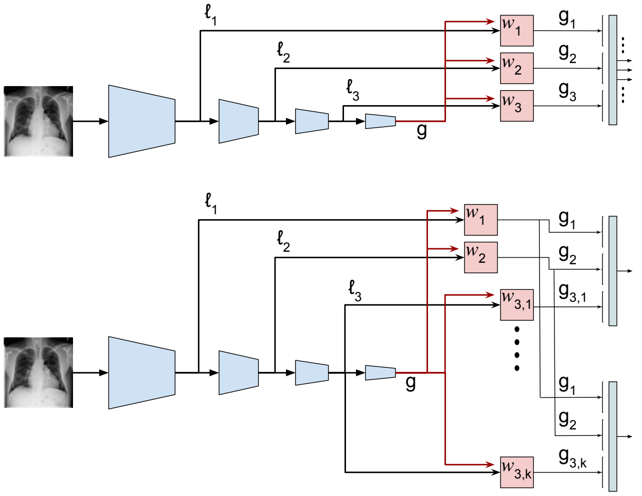


Fig. 1: Schemes of CNN models with image based attention (top) and class based attention (bottom).

To train the network we use a binary cross-entropy loss with a class weight vector since our dataset is unbalanced. The weight vector was calculated as the inverse frequency of each label. To encourage that each of the class based attentions focuses on a single pathology, we applied a regularization mechanism based on the following observation. In our training datasets we have ground truth pathology labels but we do not have an explicit ground truth localization map of a pathology. However, we do know that in case the pathology does not exist in the image it is meaningless to localize it and we therefore expect the weight matrix w_s of that class to be almost flat. Let the training image dataset be $x_1, ..., x_n$ and the binary labels associated with x_t are $y_{t1}, ..., y_{tk}$. We denote by n_i the number of positive examples from class i. We propose the following regularized classification loss:

$$\begin{aligned} L = -\sum_{t=1}^{n} \sum_{i=1}^{k} (c_i \cdot y_{ti} \log p(y_{ti} = 1 | x_t; \theta) + \\ (1 - y_{ti})(\log p(y_{ti} = 0 | x_t; \theta) + \lambda H(w_{3i}(x_t)))) \end{aligned} \tag{5}$$

where H is the distribution entropy and λ is a regularization coefficient that is tuned on the validation set. The attention distribution map $w_{3i}(x_t)$ is obtained by applying the network to the image x_t and computing the class based attention for class i. The uniform distribution is the one with maximal entropy. Therefore, by maximizing the attention map entropy in the case of no pathology in the image, we encourage the attention distribution to be uniform. To address the problem of unbalanced class sizes, the weight of a positive example from class i was set to $c_i = n_{min}^2/n_i$, such that $n_{min} = \min(n_1, ..., n_k)$.

Classes \ Model	**ResNet**	**Image Based**	**Class Based**
Atelectasis	0.76	0.78	**0.79**
Cardiomegaly	0.88	0.90	**0.91**
Consolidation	0.71	0.72	**0.76**
Edema	0.81	**0.86**	**0.86**
Effusion	0.81	**0.83**	**0.83**
Emphysema	0.91	0.92	**0.93**
Fibrosis	0.79	0.81	**0.82**
Hernia	0.89	0.91	**0.93**
Infiltration	0.70	0.71	**0.72**
Mass	0.82	**0.83**	**0.83**
Nodule	0.75	0.77	**0.79**
Pleural Thickening	0.76	0.77	**0.78**
Pneumonia	0.67	0.72	**0.73**
Pneumothorax	0.86	0.86	**0.87**
Mean AUC	0.794	0.812	**0.823**

Table 1: AUC binary classification results for the 14 thoracic pathologies in the ChestX-Ray14 dataset [11].

3. EXPERIMENTAL RESULTS

To test the performance of the class based attention architecture we trained 3 models with the same split of 3-fold cross-validation. The first model was a plain ResNet-50 [13] classifier, the second model was ResNet-50 with image based attention [2], and the third model is ResNet-50 with our class based attention. Note that the class based attention model had less than 1% more parameters than the image based model. We used the same attention block as described in detail in [2] (and replaced the backbone from VGG to ResNet50 that yields better results). We kept the training environment the same for all 3 models, e.g. optimizer and learning rate scheduling. The training of each model was stopped when no decrease in the loss was evident in the validation set.

We first evaluated our method on the ChestX-Ray14 dataset [11]. It comprises 112,120 frontal-view X-ray images of 30,805 unique patients with the text-mined fourteen disease image labels (where each image can have multi-labels), mined from the associated radiological reports using natural language processing. This dataset contains 15 classes. One of the classes is a “No-Findings” class, which we do not use. The 14 other classes, that we use in this work, contain common thoracic pathologies including Atelectasis, Consolidation, Infiltration, Pneumothorax, Edema, Emphysema, Fibrosis, Effusion, Pneumonia, Pleural-thickening, Cardiomegaly, Nodule, Mass and Hernia. The dataset was extracted from the clinical PACS database at the National Institutes of Health Clinical Center and consists of 60% of all frontal chest X-rays in the hospital. The dataset was split into training (70%), validation (10%), and test (20%) set. The spitting ensured no patient overlap between the splits.

The per-class area under the ROC curve was computed over the test set to assess the performance. The AUC statistical validity was estimated using bootstraping [14]. Table 1 shows the AUC results for each class and the average AUC. As can be seen, using standard image based attention we were able to improve on the standard ResNet-50 by $\sim 2\%$ average AUC, and by using our class based attention the performance increased by another $\sim 1\%$. The classification improvement trends were consistent for all the 14 thoracic pathologies.

Classes \ Model	**ResNet**	**Image Based**	**Class Based**
Enlarged Cardiomediastinum	0.56	0.52	**0.62**
Cardiomegaly	0.76	0.78	**0.80**
Lung Opacity	0.82	0.82	**0.84**
Lung Lesion	**0.89**	0.83	0.82
Edema	0.87	**0.88**	**0.88**
Consolidation	0.81	0.85	**0.86**
Pneumonia	0.51	0.60	**0.68**
Atelectasis	0.63	0.64	**0.69**
Pneumothorax	0.85	**0.91**	**0.91**
Pleural Effusion	0.89	0.89	**0.91**
Pleural Other	**0.99**	0.92	0.91
Mean AUC	0.780	0.788	**0.811**

Table 2: AUC binary classification results for the 11 selected thoracic pathologies of CheXpert dataset [12].

We next evaluated our model on the CheXpert [12] dataset which is a large dataset containing 224,316 chest radiographs of 65,240 patients. This dataset is comprised of 15 classes from which we use 11 common thoracic pathologies, including Enlarged Cardio Mediastinum, Cardiomegaly, Lung Opacity Lung Lesion Edema, Consolidation, Pneumonia, Atelectasis, Pneumothorax, Pleural Effusion and Pleural Other. We did not use the No-Findings, Fracture, and Support devices categories.

Table 2 shows AUC results for each class and the average AUC. As can be seen, using standard image based attention we were able to improve on the plain ResNet-50 by 0.8% average AUC. By using our class based attention, the performance increased by another 2.3% on average from the image-based model. The classification improvement trends were consistent for most of the thoracic pathologies. An improvement of 10% was seen for the Enlarged Mediastinum category. The only class where attention did not help was the "Pleural Other" category which is a mix of pathologies that can be located anywhere around the lungs. Note that most works on CheXpert dataset, e.g. [15], focus on data augmentation, modified loss functions end network ensembles which are all complementary to our network architecture.

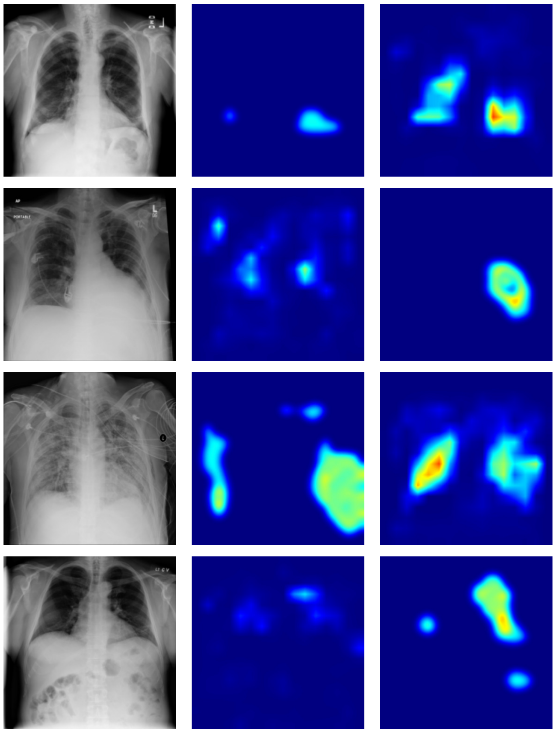
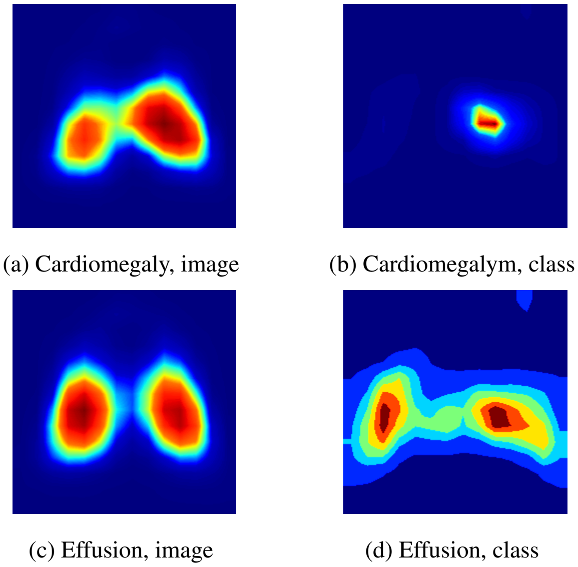


Fig. 2: Examples of image based attention (middle column) and class based attention (right column) of images from atelactasis (top), consolidation, emphysema and mass (bottom) classes.

We next show examples of the attention maps extracted from the ChestX-Ray14 dataset [11]. Fig. 2 depicts several examples of an X-ray image and its corresponding image based and class based attention maps. It shows that the class based attention is much more localized and focuses on a specific area that is relevant to the pathology, e.g. for consolidation the class based focused on the bottom of the left lung which is much more opaque due to fluids. For the Emphysema class the class based attention map focused on the inner edges of the lungs whereas the image based attention focused on unrelated areas.

We next compare the average attention masks exported from the image based and the class based attention models. Fig. 3 shows the image and class attention masks for two classes, Cardiomegaly and Effusion, out of the 14. The attention masks of Cardiomegaly were obtained by averaging the masks created from all the test images that had Cardiomegaly as the only positive class. Note that in the case of image based attention, a single mask is created for each image and in the case of class based attention the Cardiomegaly mask is one of 14 masks created in parallel for each image. The mask images of Effusion were created in a similar way. Figs. 3a and 3c show that the two image based masks are similar and they cover the entire lungs' area that is relevant to all the thoracic pathologies. In the case of Cardiomegaly, which is an enlarged heart, Fig. 3b shows that the class based mask was indeed localized at the heart area. In the case of Effusion or fluid on the lungs, Fig. 3d shows that the class based mask is indeed localized in the bottom part of the lungs which is the correct location for assessing existence of Effusion.



(a) Cardiomegaly, image

(b) Cardiomegalym, class

(c) Effusion, image

(d) Effusion, class

Fig. 3: Average attention masks for image based attention (left) and class based attention (right).

To conclude in this study we described an attention mechanism that applies a different localization procedure for each class in the task. We concentrated on automatic detection of common thorax disease categories from a chest X-ray image. An important component of our approach is a loss function that encourages the localization mask to be similar to a uniform distribution in the case where there is no pathology in the image. The class based attention concept is general and can be incorporated into other medical imaging tasks.

4. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access. Ethical approval was not required as confirmed by the license attached with the open access data.

5. REFERENCES

[1] Monty Charles, "Unscear report 2000: sources and effects of ionizing radiation," *Journal of Radiological Protection*, vol. 21, no. 1, pp. 83, 2001.

[2] Saumya Jetley, Nicholas A. Lord, Namhoon Lee, and Philip H. S. Torr, "Learn to pay attention," in *International Conference on Learning Representations (ICLR)*, 2018.

[3] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International Conference on Machine Learning (ICLR)*, 2015.

[4] Yi Wang, Zijun Deng, Xiaowei Hu, Lei Zhu, Xin Yang, Xuemiao Xu, Pheng-Ann Heng, and Dong Ni, "Deep attentional features for prostate segmentation in ultrasound," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2018.

[5] Jo Schlemper, Ozan Oktay, Michiel Schaap, Mattias Heinrich, Bernhard Kainz, Ben Glocker, and Daniel Rueckert, "Attention gated networks: Learning to leverage salient regions in medical images," *Medical Image Analysis*, vol. 53, pp. 197–207, 2019.

[6] Dong Nie, Yaozong Gao, Li Wang, and Dinggang Shen, "Asdnet: Attention based semi-supervised deep networks for medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2018.

[7] Abhijit Guha Roy, Nassir Navab, and Christian Wachinger, "Concurrent spatial and channel 'squeeze & excitation'in fully convolutional networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2018.

[8] Qingji Guan, Yaping Huang, Zhun Zhong, Zhedong Zheng, Liang Zheng, and Yi Yang, "Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification," *arXiv preprint arXiv:1801.09927*, 2018.

[9] Hongyu Wang, Haozhe Jia, Le Lu, and Yong Xia, "Thorax-net: an attention regularized deep neural network for classification of thoracic diseases on chest radiography," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 2, pp. 475–485, 2019.

[10] Zhaoyang Niu, Guoqiang Zhong, and Hui Yu, "A review on the attention mechanism of deep learning," *Neurocomputing*, vol. 452, pp. 48–62, 2021.

[11] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers, "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[12] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al., "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proceedings of the AAAI conference on artificial intelligence*, 2019, vol. 33, pp. 590–597.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[14] Honghu Liu, Gang Li, William G Cumberland, and Tongtong Wu, "Testing statistical significance of the area under a receiving operating characteristics curve for repeated measures design with bootstrapping," *Journal of Data Science*, vol. 3, no. 3, pp. 257–278, 2005.

[15] Hieu H. Pham, Tung T. Le, Dat Q. Tran, Dat T. Ngo, and Ha Q. Nguyen, "Interpreting chest X-rays via CNNs that exploit hierarchical disease dependencies and uncertainty labels," *Neurocomputing*, vol. 437, pp. 186–194, 2021.