

Supplementary for Interactive Medical Image Analysis with Concept-based Similarity Reasoning

1. Network architecture

We use ConvNext-T [3] backbone with ImageNet [1] pretrained as the feature extractor F . The concept head C consists of a 1×1 convolution layer without bias, a global average pooling layer followed by a Sigmoid activation. The projector P includes a 3 blocks of 1D convolution layer, followed by an IBN [4] and a ReLU activation with a residual connection. The final output from P is L_2 -normalized. The task head H is a single Linear layer. We set the number of cluster prototypes $M = 100$ for each concept.

2. Train time interaction

We present the procedure of train-time interaction in Algorithm 1.

Algorithm 1 Train-time interaction

Input: Learned concept prototypes $\{p^{k_m}\}$ $\triangleright k$ is the concept index, m is the instance index..
for each concept prototype p^{k_m} **do**
 Retrieve the interpretable version $\mathcal{I}(p^{k_m})$ of the concept prototype.
 Visualize p^{k_m} by overlaying the similarity map S^{k_m} of p^{k_m} on $\mathcal{I}(p^{k_m})$.
 Present the visualized concept prototype p^{k_m} to the radiologist.
 Radiologist review the visualized p^{k_m}
 if p^{k_m} is a valid signal **then**
 Retain the prototype
 else
 Discard the prototype from $\{p\}$
 end if
end for
Update the set of valid prototypes $\{p\}$ by removing discarded prototypes

3. Training configurations

For the multi-prototypes contrastive loss, we set $\lambda = 20, \gamma = 1000, \delta = 0.1$ in all 3 datasets.

3.1. TBX11K

The TBX11K dataset contains 11k CXRs for the task of Tuberculosis, Sick but not Tuberculosis, and Healthy classification. It contains image-level annotations of the 3 mentioned classes. Besides, there are 400 of images having box-level annotations for the Tuberculosis class. We train with the original image size of 512×512 . Due to the relatively small size of the dataset, we leveraged Chexpert [2] to pretrain the baselines. We also utilized the set of findings in Chexpert as concepts to supports the 3 target classes in the TBX11K dataset for CSR and CBM. There are 15 findings: No Finding, Support Devices, Enlarged Cardiomegaly, Cardiomegaly, Fracture, Lung Opacity, Edema, Consolidation, Pneumonia, Lesion, Atelectasis, Lesion, Pneumothorax, Pleural Effusion and Pleural Other.

Because the No Finding concept represents the absence of findings, we cannot take the CAM like we do with other concepts. To create the local concept vector v^{NF} for No Finding, we argue that a Normal image can be exploited at every region because they are all normal. To this end, we divide the feature map \mathbf{f} spatially by a 2×2 grid and take the sum of each grid to obtain 4 v^{NF} local concept vectors from each normal image. However, we just leverage them for training P , but not using the No Finding prototypes in for Similarity maps calculation, resulting explanation size of 14 for CSR and CBM instead of 15 of 15 concepts.

3.2. VinDr-CXR

The VinDr-CXR dataset contains 18k CXRs for multi-label classification. It has 6 image-level (global) annotations, and 22 box-level (local) annotations. We define the local annotations as the concepts and the global annotations as the target classes. We filter out low-numbered classes, obtaining the final list of 15 concepts: Aortic enlargement, Atelectasis, Cardiomegaly, Calcification, Consolidation, ILD, Infiltration, Lung Opacity, Mediastinal shift, Nodule/Mass, Pulmonary fibrosis, Pneumothorax, Pleural thickening, Pleural effusion, Rib fracture. For the targets, we choose 3 classes: Lung tumor, Pneumonia and Tuberculosis. We train with the original image size of 512×512 . We also pretrain the baseline with Chexpert. For CSR and CBM, we continue to train the Concept model to predict

the local annotations. Notice that we do not use the bounding boxes of the local annotations but only treating them as image-level annotations for training the concepts.

3.3. ISIC

The ISIC 2017 dataset contains 3K dermoscopic images for skin lesions analysis. It comprises of 3 tasks: (1) Lesion segmentation; (2) Localization of visual dermoscopic features; (3) Disease classification. We use the visual features annotations of task 2 as the concepts and the disease annotations of task 3 as the target classes. There are 4 concepts: milia like cyst, pigment network, negative network and streaks. The 3 targets include: seborrheic keratosis, nevus/melanoma, no findings. We pretrain the baselines with the dataset from the 2019 version of the competition which contains 35k dermoscopic images. In this dataset, we use the image size of 224×224 .

4. Prototype projection

Proto-part methods involve a step of prototype projection, where the prototypes are replaced by the nearest latent patch from the actual training image. In contrast, in CSR, a concept prototype represents a broader concept rather than a specific latent training patch. Replacing a concept prototype with the nearest local concept vector can limit its representational capacity. In fact, we observed a 0.7% drop in F1-score on the TBX11K dataset when applying prototype projection. Additionally, concept prototypes in CSR are used to highlight relevant concepts in the image through comparison, making CSR interpretable from a visual grounding and explanation perspective.

5. Pointing Game

The qualitative result of the pointing game is illustrated in Fig.1.

References

- [1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 1
- [2] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silvana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghighi, 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*, volume 33, pages 590–597, 2019. 1
- [3] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. 1
- [4] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the european conference on computer vision (ECCV)*, pages 464–479, 2018. 1

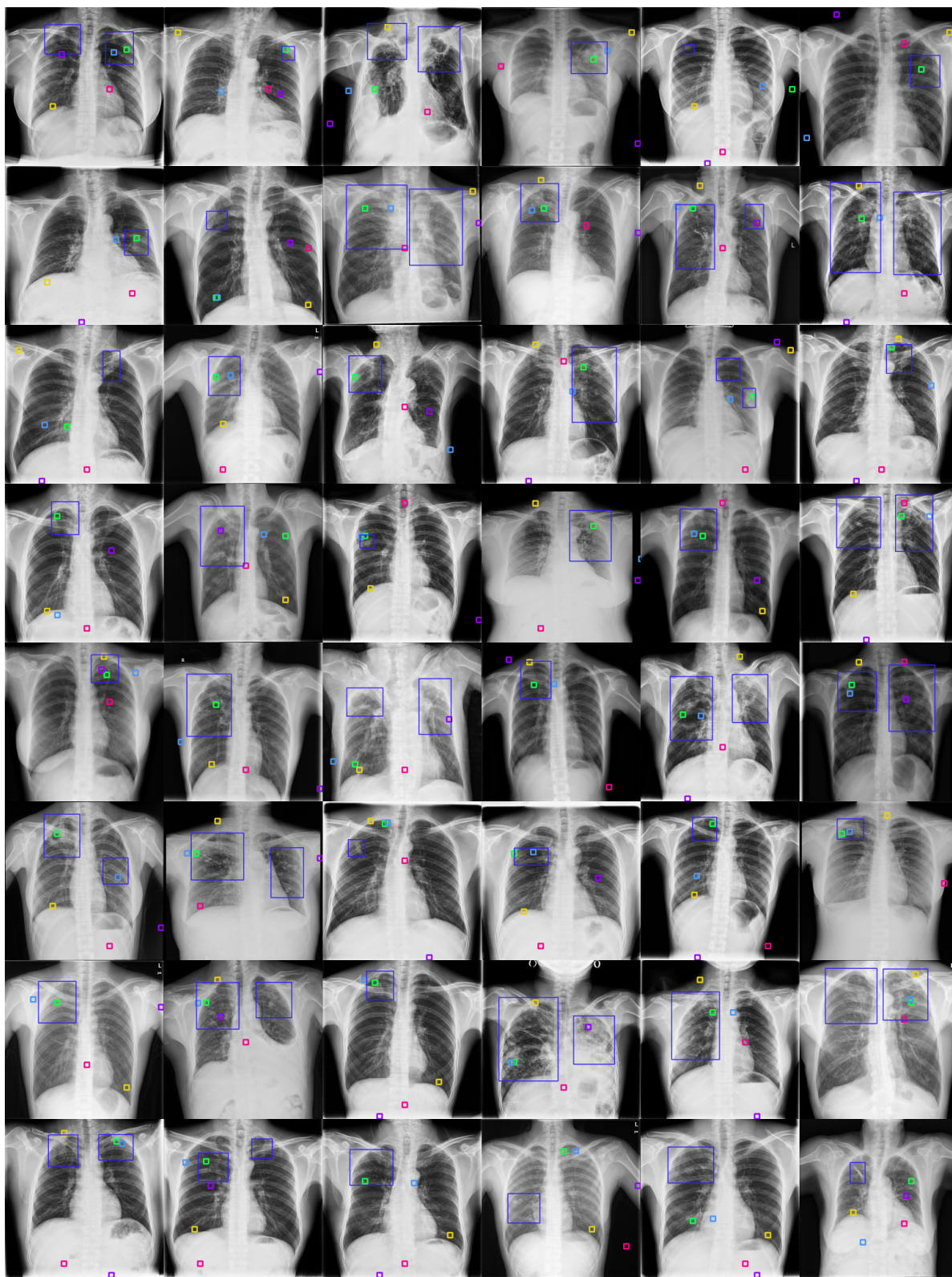


Figure 1. Qualitative result of the Pointing Game. We visualize the maximum activation point of each method as the square on to a subset of the TBX11K dataset. ■:CSR, ■:CBM, ■:PIP-Net, ■:ProtoPNet, ■:ProtoTree